Supermassive Black Holes from Gravothermal Collapse of Fractional Self-Interacting Dark Matter halos

Jason Pollack
Advisor: Paul Steinhardt

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts
Department of Physics
Princeton University

May 2012

This paper represents my own work in accordance with University regulations.

_____

May 7th, 2012

Abstract

The "standard model of cosmology", abbreviated as ΛCDM, postulates a form of cold dark matter which is weakly interacting both with itself and with baryonic matter. Self-interacting dark matter (SIDM) is a proposed modification to ΛCDM cosmology which makes the dark matter self-interacting with a non-negligible scattering cross section. The effect is to reduce clumping on small scales, which could help to resolve the so-called "missing satellites problem" and "core/cusp problem": disagreements between ΛCDM-based simulations and observations on galactic or subgalactic scales. In this paper, we use SIDM to solve a different problem: growing billion-solar mass black holes in the center of galaxies at high redshifts, as required by observations of high-redshift quasars. We consider halos in which a small fraction of the dark matter has non-negligible self-interaction, and show that gravothermal collapse of these fractional SIDM halos is able to produce intermediate-mass black holes without the need for star formation or gas accretion, alleviating tensions within the standard ΛCDM picture.

Acknowledgements

In the process of working on my thesis I consulted with many active researchers in the field, both at Princeton and elsewhere. In particular, I would like to thank Renyue Cen and Naoki Yoshida for useful in-person discussions, on the missing satelites problem and cusp problem with the former and on his simulations of SIDM clusters and the general concept of growing black holes from gravothermal collapse with the latter. In addition, I exchanged useful emails with Stu Shapiro and Shmulik Balberg, as well as Jun Koda and Paul Shapiro, on their respective papers on simulating gravothermal collapse of SIDM halos. Thanks especially to Jun for his willingness to give in-depth explanations of the implementation aspect of the problem. Finally, I had many useful meetings with David Spergel, my second reader: I am particularly grateful for his guidance in critically evaluating the astrophysical constraints on SIDM, and, especially, for directing me to a useful approximation for the halo mass function.

Of course, all of these contributions pale next to the constant guidance and encouragement I have received from my advisor, Paul Steinhardt, not just while working on my thesis but over the entire span of my undergraduate career at Princeton. I would never have managed to survive this year, especially, without his support, both academically and generally as a mentor. I look forward to many more collaborations with him as my career as a physicist continues.

In addition, I thank Donald Knuth and Leslie Lamport for creating TEX and LATEX, respectively, without which this thesis would not have been possible in the most literal sense, as well the creators of LYX, without whom the writing of this thesis would still have been possible but would have enacted an even more crushing toll on my sanity.

More generally, I need to thank several people without whom my experience at Princeton would have been significantly poorer. Thank you to Lyman Page, Suzanne Staggs, Joe Fowler, Jon Sievers, and Lucas Parker for allowing me to work with you on ACT, ABS, and everything in between. Thank you to Cody Burton, Yuliya Dovzhenko, and Anasua Chatterjee for helping me survive all those problem sets, and thanks again to Cody for keeping me sane in general, especially last year and this year.

Finally, I need to thank my parents, without who I would of course not exist, and who

have encouraged me to do something insane like becoming a physicist rather than going into finance, even if they might occasionally prefer to retract some of that encouragement.

Contents

1 Introduction

The standard model of cosmology (abbreviated as ΛCDM) postulates that a cosmological constant Λ exerting negative pressure and nearly collisionless cold dark matter, together with baryonic matter and radiation (including neutrinos), comprise the energy density of the universe. Given the homogeneous and isotropic nearly scale-free Gaussian spectrum of initial perturbations observed in the cosmic microwave background [1], ΛCDM makes detailed predictions for the subsequent structure formation history of the universe [2, 3]. On large scales these predictions have been spectacularly confirmed by observations, but a number of potential problems for ΛCDM at and below the scale of individual galaxies have been identified, such as a discrepancy between the number of Milky Way satellites and the number of dark matter subhalos predicted by simulations [4, 5, 6] and observations of dwarf galaxies with flat rather than divergent central density profiles [7, 8]. In addition, one of the most natural explanations for the current dark matter abundance, the hypothesis that the CDM consists of thermally-produced weakly interacting massive particles (WIMPs), has been placed under severe tension by the continuing failure of experiments to find such particles [9].

This state of affairs provides a motivation to consider a more complex dark matter sector than one comprised entirely of monolithic collisionless cold dark matter. Given the known complexity of the Standard Model (SM) of particle physics, which is only concerned with the baryonic matter making up less than 5% of the energy density of the universe, it is not unnatural to postulate that dark matter, which contributes four times as much to the energy density, is at least as complicated. Of course, rather than immediately moving to this level of complexity, it is sensible to add complicating factors (or, equivalently, new free parameters) one at a time and see what new phenomena result.

One such complicating factor is the possibility that some or all of the dark matter could have a non-negligible self-scattering cross section, rather than the weak-scale cross section usually considered. This "self-interacting dark matter" (SIDM) scenario was first proposed by Spergel and Steinhardt [10] as a mechanism for smoothing out substructure on small scales while simultaneously avoiding constraints on free-streaming particles like warm or hot dark matter. Observations subsequently placed severe constraints on allowed scattering

cross sections [11, 12, 13], although more complicated scenarios with velocity-dependent cross sections may evade these bounds [14, 15].

In this paper, we instead consider SIDM as a mechanism for generating seed black holes in the center of galaxies at high redshifts via gravothermal collapse [16, 17]. Observations of high-redshift quasars [18, 19, 20] indicate the presence of billion-solar-mass supermassive black holes (SMBHs) in the center of galaxies at $z \gtrsim 6$; simulations of galaxy formation and mergers [21] indicate that this is possible in the ΛCDM picture, but only if optimistic assumptions about seed black hole formation, SMBH accretion rates, and black hole mergers are made. The recent observation of a $z = 7.085$ quasar hosting a black hole with mass $2 \times 10^9 M_\odot$ [22] is even harder to reconcile with the standard picture. We show that a self-interacting component of the dark matter making up a small fraction ($f \approx 10^{-2}$) of the total mass of a galactic halo can rapidly undergo gravothermal collapse to form central seed black holes, which can then easily accrete baryons to reproduce the existing abundance of high-redshift SMBHs. Because most of the dark matter in halos remains collisionless, the SIDM component may have a large cross section while simultaneously evading existing constraints from observations of galaxies and clusters.

The remainder of this paper is laid out as follows. Section 2 reviews the challenges to the ΛCDM paradigm from small-scale observations. Section 3 introduces self-interacting dark matter as a means of explaining these observations and reviews the models which have been proposed. In Section 4, we summarize existing constraints on SIDM and re-evaluate them in light of more recent astrophysical data. Section 5 reviews SMBH formation and evolution in the standard ΛCDM picture, as well as proposed mechanisms of seed black hole formation. Section 6 explains the mechanism of gravothermal collapse and applies the gravothermal fluid approximation [23, 24] to the core collapse of a fractional SIDM halo. Section 7 presents an alternative scenario of SMBH formation and evolution using SIDM cores as black hole seeds. Finally, we summarize and conclude in Section 8.

2 Motivation

2.1 The ΛCDM Paradigm

The simplest ΛCDM model describes a flat universe consisting only of photons and three additional relativistic species (the nearly massive neutrinos), a cosmological constant Λ with negative pressure (equation of state $w = -1$), collisionless cold dark matter, and baryons. It can be described using six parameters, typically taken to be the following:

$$\left\{100\Omega_b h^2,\ \Omega_c h^2,\ \Omega_\Lambda,\ n_s,\ \tau,\ \Delta^2_{\mathcal{R}}(k_0)\right\}. \tag{1}$$

Here $\Omega_{b,c,\Lambda}$ are the current densities of baryons, cold dark matter, and the cosmological constant, respectively, assuming $\Omega_b + \Omega_c + \Omega_\Lambda = 1$ (so the density in photons and neutrinos can be neglected), $100h$ km s^{-1} Mpc^{-1} is the current value of the Hubble constant, the primordial spectrum of scalar fluctuations is assumed to be purely adiabatic, homogeneous, and isotropic with a spectral index of n_s and an amplitude of $\Delta^2_{\mathcal{R}}(k_0)$ at $k_0 = 0.002$ Mpc^{-1}, and the optical depth τ of the reionized plasma created by ionization of neutral hydrogen by ultraviolet light from the first generation of stars gives a probability $1 - e^{-\tau}$ that a given photon in the cosmic microwave background (CMB) reaching us today has been scattered.

The six parameters can be measured most directly by CMB experiments such as WMAP [1] and ACT [25], which derive them from the angular power spectrum of CMB temperature anisotropies (given a prior measurement of $T_{\rm cmb} \approx 2.725$ K [26, 27, 28]). In the simplest approximation [29], $\Omega_b h^2$ and $\Omega_c h^2$ can be found from the *ratios of heights* of successive acoustic peaks in the power spectrum; given these values and the assumption of zero curvature, the *angular scales* of peaks give a measurement of h^2 and thus Ω_Λ. Finally, the overall magnitude of the spectrum, given by any one *peak height*, is a function of τ and $\Delta^2_{\mathcal{R}}(k_0)$; τ can be measured independently by observations of CMB *polarization*, allowing the determination of $\Delta^2_{\mathcal{R}}(k_0)$ as well. WMAP 7-year results [1] alone yield the following

values (68% CL):

$$\begin{aligned} 100\Omega_b h^2 &= 2.249^{+0.056}_{-0.057} \\ \Omega_c h^2 &= 0.1120 \pm 0.0056 \\ \Omega_\Lambda &= 0.727^{+0.030}_{-0.029} \\ n_s &= 0.967 \pm 0.014 \\ \tau &= 0.088 \pm 0.015 \\ \Delta^2_{\mathcal{R}} &= (2.43 \pm 0.11) \times 10^{-9}. \end{aligned} \tag{2}$$

The six-parameter model with these parameters provides an excellent fit to CMB data, as well as to data from a wide range of other astrophysical measurements, including big bang nucleosynthesis (BBN) [30, 31], distance measurements from Cepheid variables [32, 33] and supernovae [34, 35], measurements of large scale structure from baryon acoustic oscillations (BAO) [36, 37] and galaxy clusters [38, 39], and strong [40, 41] and weak [42, 43] lensing measurements. No conclusive evidence has thus far been found for any of many proposed extensions to the six-parameter model, such as nonzero curvature, primordial tensors (gravitational waves), running of the scalar spectral index, additional relativistic species (warm or hot dark matter), nonadiabaticity (isocurvature fluctuations), a dark energy equation of state which is non-constant or with $w \neq -1$ (quintessence models), parity violation, cosmic strings, etc.

Any alteration to this picture which has non-negligible observational effects, then, must satisfy two general constraints if it is to fit that data. First, it must not make large changes *at early times*, to avoid the stringent constraints from BBN and the CMB. Second, it must avoid significant changes *on large scales*, in light of the excellent agreement between structure formation simulations and large-scale surveys. One major class of changes obeys these constraints by altering the primordial power at small scales; examples include warm dark matter [44] and direct suppression of small-scale power in certain inflationary models [45]. However, there are strong constraints from the Lyman-α forest [46] on the small-scale power, and we do not consider such models further in this paper. The remaining class of models alters the properties of the ingredients already present in ΛCDM—for example, by postulating a dark sector rather than one monolithic dark matter particle. This will be our approach. To motivate it, we first consider the standard picture, where the cold dark matter is a weakly interacting massive particle (WIMP), then review the possible inadequacies of

this picture.

2.2 The "WIMP Miracle"

If there exists a conserved quantum number which is not carried by any of the particles in the Standard Model (such as R-parity, in the supersymmetric case), the lightest particle χ which carries this quantum number will be stable against decay: particles can only be destroyed by pair annihilation at a rate per unit volume $n_\chi^2 \langle \sigma_\chi v \rangle$, where n_χ is the number density of χ particles (as well as of $\bar{\chi}$ particles, if χ is not its own antiparticle) and σ_χ is its annihilation cross section[1]. Following [47], consider a comoving volume a^3 in the early universe, when all lighter particles are in chemical and thermal equilibrium. Then the evolution of n_χ is described by a Boltzmann equation,

$$\frac{d\left(n_\chi a^3\right)}{dt} = -\left\langle \sigma_\chi v \right\rangle \left(n_\chi^2 - n_{\chi,eq}^2\right) a^3, \tag{3}$$

or, writing $H = \frac{\dot{a}}{a}$,

$$\frac{dn_\chi}{dt} = -3Hn_\chi - \left\langle \sigma_\chi v \right\rangle \left(n_\chi^2 - n_{\chi,eq}^2\right), \tag{4}$$

where $n_{\chi,eq}$ is the equilibrium number density, so that the creation rate per unit volume is $n_{\chi,eq}^2 \langle \sigma_\chi v \rangle$ to balance the annihilation rate in equilibrium. At early times, where $T \gg m_\chi$, we must have $n_\chi \propto T^3$ on dimensional grounds, while $H \propto T^2$ in the radiation-dominated epoch, as is shown in (15) below, so the expansion term $3Hn_\chi$ is negligible at early times but dominant at late ones. As the temperature falls below m_χ, $n_{\chi,eq}$ is exponentially suppressed and becomes negligible. The two remaining terms on the right-hand side become equal to each other at *freezeout*, where

$$n_{\chi,f} \left\langle \sigma_\chi v_f \right\rangle = 3H_f. \tag{5}$$

For weak-scale cross sections, the freezeout temperature is [48]

$$T_f \approx m_\chi/20; \tag{6}$$

[1] In this section, we set $\hbar = c = k_B = 1$ in all intermediate steps for clarity.

beyond this point the χ particles are too diluted to annihilate, and their comoving number density remains constant. In the standard picture, the entropy per comoving volume is also constant, so we can relate the χ density at freezeout to its current density, $n_{\chi,0}$:

$$\frac{n_{\chi,0}}{s_0} = \frac{n_{\chi,f}}{s_f}, \tag{7}$$

where s denotes the entropy density, and the current entropy density is known to be [49]

$$2889.2 \text{ cm}^{-3} \tag{8}$$

The fraction of the total energy density in χ particles is given by $\Omega_\chi = m_\chi n_{\chi,0}/\rho_{c,0}$, where the critical density is given by [49]

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} \approx 1.054 \cdot 10^{-5} h^2 \text{ GeV} \cdot \text{cm}^{-3}, \tag{9}$$

so, using (7) and (5),

$$\Omega_\chi = \frac{3 m_\chi s_0 H_f}{s_f \rho_c \langle \sigma_\chi v_f \rangle}. \tag{10}$$

It remains to find expressions for H_f and s_f as functions of T. First, recall the first Friedmann equation,

$$H^2 = \frac{8\pi G}{3}\rho, \tag{11}$$

where we have neglected curvature and the cosmological constant as irrelevant in the early universe. In the radiation-dominated epoch, the density is given by the sum of contributions of the various relativistic species ρ_i, with

$$\rho_i = \int_0^\infty n_i\left(p,T\right) dp\sqrt{p^2+m_i^2} = \int_0^\infty \frac{4\pi g_i p^2}{\left(2\pi\right)^3} \frac{dp\sqrt{p^2+m_i^2}}{\exp\left(\sqrt{p^2+m_i^2}/T\right) \pm 1}, \tag{12}$$

where g_i is the number of helicity states of the ith particle and the sign is positive (negative) for fermions (bosons). For relativistic particles, $p^2 \gg m_i^2$, the integral can be solved exactly

to give

$$\rho_i = \frac{4\pi g_i}{(2\pi)^3} \begin{cases} \frac{\pi^4 T^4}{15} & \text{bosons} \\ \frac{7\pi^4 T^4}{120} & \text{fermions} \end{cases}. \tag{13}$$

Then, collecting the helicity states together as

$$g_* = \sum_{\text{bosons}} g_i + \sum_{\text{fermions}} \frac{7}{8} g_i \tag{14}$$

gives

$$H_f = \sqrt{\frac{8\pi G}{3} \frac{4\pi g_*}{(2\pi)^3} \frac{\pi^4 T_f^4}{15}} = \sqrt{\frac{4\pi^3}{45}} \frac{g_*^{1/2} T_f^2}{m_{\text{Pl}}} \approx 1.66 \frac{g_*^{1/2} T_f^2}{m_{\text{Pl}}}, \tag{15}$$

where we have used the relation $G = 1/m_{\text{Pl}}^2$.

To find the entropy density at freezeout, write the second law of thermodynamics:

$$TdS = dU + pdV \rightarrow d(sV) = \frac{1}{T}\left[d(\rho V) + pdV\right]. \tag{16}$$

Equating the coefficients of dV and gives an expression for s,

$$sdV = \frac{1}{T}(\rho + p)dV \rightarrow s = \frac{\rho + p}{T} = \frac{\frac{4}{3}\rho}{T}, \tag{17}$$

since $w = p/\rho = 1/3$ for relativistic species. Inserting (13) gives

$$s_f = \frac{2\pi}{45} g_* T^3 \approx 0.44 g_* T_f^3. \tag{18}$$

Inserting (15) and (18) into the definition of Ω_χ (10), and recalling (6) $T_f \approx m_\chi/20$ gives

$$\Omega_\chi \approx \frac{3 m_\chi s_0}{\rho_{c,0} \langle \sigma_\chi v_f \rangle} \cdot 1.66 \frac{g_*^{1/2} m_\chi^2}{20^2 m_{\text{Pl}}} \cdot \frac{20^3}{0.44 g_* m_\chi^3} = \frac{22.636 \hbar^2 c^{-1}}{m_{\text{Pl}} \langle \sigma_\chi v_f \rangle} \cdot \frac{s_0}{\rho_{c,0}}, \tag{19}$$

where we have inserted factors of $\hbar$ and c to make the expression dimensionless. Evidently the relic abundance depends only on the annihilation cross section, not independently on the mass. Assuming that the total number of light degrees of freedom $g_* \approx 100$, valid for

temperatures in the 1–1000 GeV range [48], and inserting numerical values yields

$$\begin{aligned}\Omega_\chi h^2 &\approx \frac{22.636 \cdot \left(2889.2 \text{ cm}^{-3}\right) \hbar^2 c^{-1}}{\left(1.221 \cdot 10^{19} \text{ GeV}\right)\left(1.054 \cdot 10^{-5} \text{ GeV} \cdot \text{cm}^{-3}\right)} \frac{1}{\langle \sigma_\chi v_f \rangle} \\ &\approx \frac{5.93 \cdot 10^{-27} \text{ cm}^3 \cdot \text{s}^{-1}}{\langle \sigma_\chi v_f \rangle}. \end{aligned} \tag{20}$$

If χ is a WIMP, its low-energy annihilation cross section should be set by the weak scale, $m_w \sim 246$ GeV: to lowest order, we expect

$$\langle \sigma_\chi v_f \rangle \sim \frac{\hbar^2 g^2}{m_w^2 c} \approx g^2 \cdot 1.9 \cdot 10^{-22} \text{ cm}^3 \cdot \text{s}^{-1}, \tag{21}$$

where for convenience we have absorbed factors of 2 and π into the coupling constant g. Then the WIMP relic abundance matches the observed value of $\Omega_c h^2$ (2) when

$$0.112 \approx \frac{3.07 \cdot 10^{-5}}{g^2} \rightarrow g \approx 0.0166, \tag{22}$$

a natural value for an electroweak coupling constant. This is the "WIMP miracle"—a stable particle with a typical weak-scale cross section naturally yields a relic abundance of the same magnitude as the observed abundance of dark matter. Note that the WIMP miracle truly is "miraculous", in that it relies on a coincidence between (presumably) unchanging particle physics parameters and current cosmological ones:

$$\frac{m_w^2}{g^2 m_{pl}} \sim \frac{\Omega_\chi \rho_{c,0}}{s_0}. \tag{23}$$

2.3 CDM Challenges

2.3.1 Experimental Searches

If some part of the dark matter is made up of weakly interacting particles, it should be possible to detect their scattering off of nucleons, for example in underground detectors [50]. If the scattering is primarily coherent and spin-independent, the differential event rate (typically measured in events/kg $\cdot$ day) is the convolution of the WIMP-nucleus cross section

with the WIMP velocity distribution [51]

$$\frac{dR}{dE_R} = A^2 N_T \frac{\xi \rho_0}{m_\chi} \frac{m_N}{2\mu_1^2} \sigma^{nuc} F^2 (E_R) \int_{v \geq v_{min}} \frac{f_\oplus (\vec{v})\, d^3 v}{v}, \tag{24}$$

where E_R is the nuclear recoil energy, A is the number of nucleons per nucleus, N_T is the number of target nuclei per unit mass, ξ is the fraction of dark matter made up of WIMPs,ρ_0 is the local dark matter density, measured to be [52] $\rho_0 = 0.39 \pm 0.03$ GeV $\cdot$ cm^{-3}, m_χ is the WIMP mass, m_N the mass of the nucleus, μ_1 the WIMP-nucleon reduced mass, σ^{nuc} the WIMP-nucleon coherent scattering cross section, and $F(E_R)$ is the nuclear form factor, which determines the likelihood of the nucleus scattering with energy E_R. Finally, $f_\oplus(\vec{v})$ is the WIMP velocity distribution in the earth's rest frame, typically taken to be a Maxwellian distribution peaking around 220 km $\cdot$ s^{-1} truncated at the galactic escape velocity [53] $v_{esc} \approx 544$ km $\cdot$ s^{-1}, and v_{min} is the minimum WIMP velocity for which a recoil energy of E_R is allowed:

$$v_{min} = \sqrt{\frac{m_N E_R}{2\mu_2^2}}, \tag{25}$$

where μ_2 is the WIMP-nucleus reduced mass. Hence, given a knowledge of the background and nuclear form factor, (24) allows limits to be set in the m_χ–σ^{nuc} plane. Separation of WIMP-nucleon recoils from background can be achieved by exploiting the relative motion of the earth with respect to the galactic halo, which should result in an annual modulation of the signal as the location of the high-velocity cutoff of $f_\oplus(\vec{v})$ in the Earth's reference frame grows and shrinks [54]. The background from solar neutrinos will also show an annual modulation, but one peaking in January, when the Earth is closest to the sun, rather than in May, when the earth's velocity is parallel to the direction of the Sun's orbit around the center of the galaxy. Evidence for such a signal is inconclusive: the DAMA/LIBRA [55] and CoGeNT [56] collaborations claim to see such a signal, but data from the CDMS collaboration [57] disfavors the CoGeNT modulation at $> 98\%$ confidence.

Naively, we should expect the WIMP-nucleon scattering cross section, $\sigma^{nuc} = \sigma(\chi n \to \chi n)$, to be identical to the cross section for χ annihilation into nucleons, $\sigma(\chi\chi \to nn)$, which should make up a substantial fraction of the total annihilation cross section σ^{ann}. If the χ

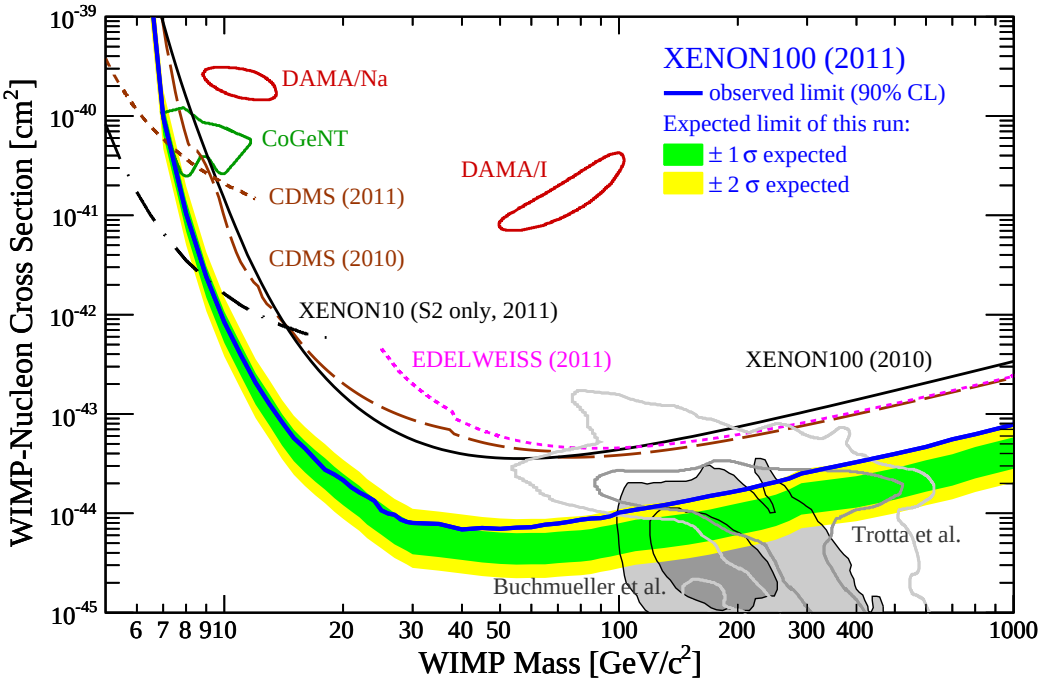

Figure 1: Exclusions and favored areas in the (spin-independent)WIMP–Nucleon cross section vs. WIMP mass plane, assuming that the dark matter abundance comes entirely from a relic WIMP. The plot, reproduced from [9], shows the observed and expected 95% confidence limits from 100 days of data taking by the XENON100 experiment (2011), as well as additional limits from earlier (2010) XENON100 data [58], EDELWEISS (2011) [59], CDMS (2010) [60], a 2011 CDMS low-mass WIMP search [61], XENON10 (2011) [62], and 90% CL favored areas from CoGeNT [63] and DAMA (without channeling) [64]. Grey contours [65] and shaded regions [66] are 1σ and 2σ preferred regions from the MSSM constrained by precision electroweak observables and collider supersymmetry searches, from 2008 and 2011 respectively.

relic abundance accounts for all the dark matter, (20) gives[2]

$$\sigma^{nuc} \lesssim 1.74 \cdot 10^{-36} \text{ cm}^2 = 1.74 \text{ pb}. \tag{26}$$

However, a number of dark matter experiments in the last several years have excluded this cross section by many orders of magnitude for WIMPs in the 5–1000 GeV range [67]. Figure 1, taken from [9], shows the limits in the m_χ–σ^{nuc} plane, assuming that the χ–nucleon interaction is spin-independent and that observed dark matter abundance comes entirely from the χ relic abundance (i.e. $\xi = 1$). For WIMP masses between 5 and 1000 GeV, σ^{nuc} is more than four orders of magnitude below the favored value; above 10 GeV, it is disfavored by more than seven.

Of course, supersymmetric models which predict that the lightest supersymmetric particle has a cross section this low still exist—the gray contours and shaded areas in Figure 1 show predicted values, taking into account existing constraints from precision electroweak data and collider searches for supersymmetry. However, these models must necessarily be heavily fine-tuned to prevent falsification from the existing exclusions. But the chief virtue of the WIMP miracle is its naturalness—the relic abundance of a *generic* WIMP matches the observed dark matter density, without the need for fine-tuning. To the extent that direct detection experiments push the WIMP parameters away from their natural values, then, we can say that the WIMP miracle is dead, or at least significantly less miraculous. Given that these natural values have been experimentally excluded, it becomes more plausible to consider more complicated models, such as ones that include a dark sector.

2.3.2 The Missing Satellites Problem

In the late 1990s, practical numerical simulations of structure formations became sophisticated enough to follow dark matter subhalos within a larger halo, allowing, for example the study of satellite subhalos within a galactic halo or galactic subhalos within the halo of a cluster. Earlier simulations showed substructure disappeared within larger halos, the "overmerging" problem [68, 69], but analytical work suggested the problem was due to numerical effects resulting from lack of resolution [70]. New simulations with greater resolution but us-

[2]Since the scattering is taking place in the low energy limit, the χ rest energy dominates the kinetic energy, so $\sigma \approx \langle \sigma_\chi v_f \rangle / c$.

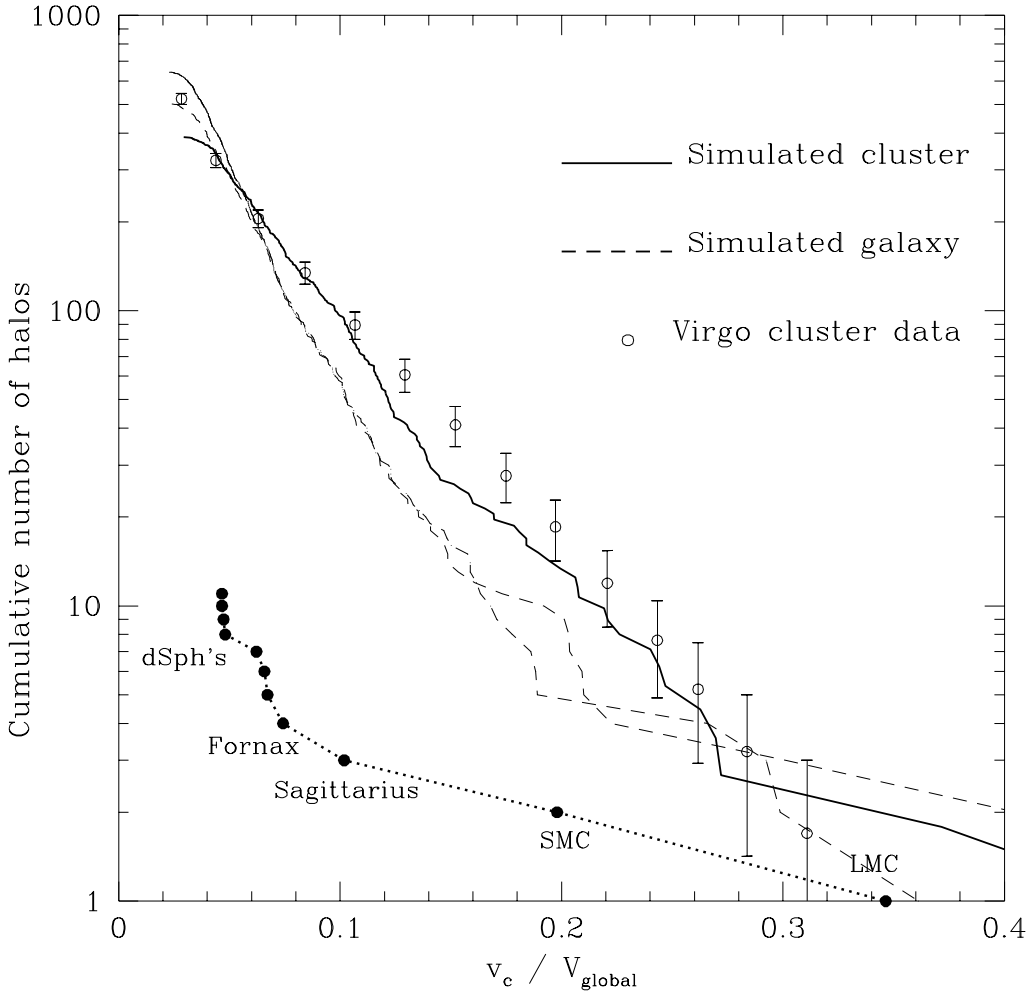

Figure 2: Cumulative number of subhalos with circular velocity v_c (normalized by the velocity $v_{\rm global}$ of the parent halo) within actual and simulated structures. The dotted curve shows Milky Way satellites observed as of 1998 [74], while the dashed lines show abundance of subhalos within a simulated galaxy of mass $2 \times 10^{12} M_\odot$, in the present day and 4 billion years earlier. At the scales of the smallest observed Milky Way satellites, the dwarf spheroidal galaxies, there is an abundance deficit of nearly two orders of magnitude. The open circles, which show the abundance of galactic subhalos within the Virgo cluster [75], match the abundances within a simulated cluster of mass $5 \times 10^{14}\ M_\odot$. Evidently CDM simulations predict roughtly self-similar substructure, but observations within galactic halos diverge from this pattern. Reproduced from [5].

ing the same theoretical models (e.g. without taking gas dynamics into account) succeeded in simulating the survival of galactic halos within clusters [71, 72, 73]. Several groups [4, 5] then simulated substructure within galaxy-scale halos and compared the results to observations of dwarf galaxy satellites within the Local Group [74]. Figure 2, taken from [5], plots the simulated cumulative abundance of substructure within galaxy- and cluster-scale halos compared to observations of the Milky Way and Virgo Cluster. CDM simulations predict that substructure should be broadly self-similar: the normalized abundance curves for the simulated cluster and galaxy are essentially identical. The cluster data is in agreement with simulations, but the number of Milky Way satellites larger than dwarf spheroidal galaxies (dSphs) is overpredicted by a factor of 50, the so-called "missing satellites problem."

One resolution for this problem is to postulate that the missing dSphs actually exist but have not yet been observed. Beginning in 2004 [76], the Sloan Digital Sky Survey has doubled the number of known dSphs while observing only $\sim 1/5$ of the sky. Correcting the observed Milky Way satellite luminosity function for luminosity bias suggests there may be ~ 500 total satellites brighter than the faintest known dSphs [77], which would entirely alleviate the missing satellites problem as originally posed. However, the problem is exacerbated on even smaller scales: recent galactic-scale simulations [78, 79] indicate that tens or hundreds of thousands of satellites with masses larger than $10^5 M_\odot$ should exist within the Milky Way.

If the actual number of dSphs is indeed overpredicted by ΛCDM, it is possible that baryonic physics could be to blame: perhaps the predicted subhalos exist, but only a fraction of them are populated by galaxies. If this is the case, we would expect that galaxy luminosity is linked either to the present-day subhalo mass [80, 81, 82], if star formation activity is linked to halo mass or concentration, or to the subhalo mass at an earlier epoch such as infall onto the galactic halo or reionization [83, 84, 85], if satellite galaxies only form at all in the largest or most concentrated halos. However, substructure data from the Aquarius simulations [6] appears to falsify such theories in the standard ΛCDM picture. Figure 3 shows the numerically derived circular velocity curves of dSph host candidates in the Aquarius E halo ($M_{\rm vir} = 1.4 \times 10^{12} M_\odot$), along with observed circular velocity measurements from each of the nine bright Milky Way dSphs currently in equilibrium. If galaxy formation is dependent on subhalo properties, we should certainly expect that the subhalos with the largest mass, or equivalently the largest maximum circular velocity, at a given epoch should host dSphs. Yet

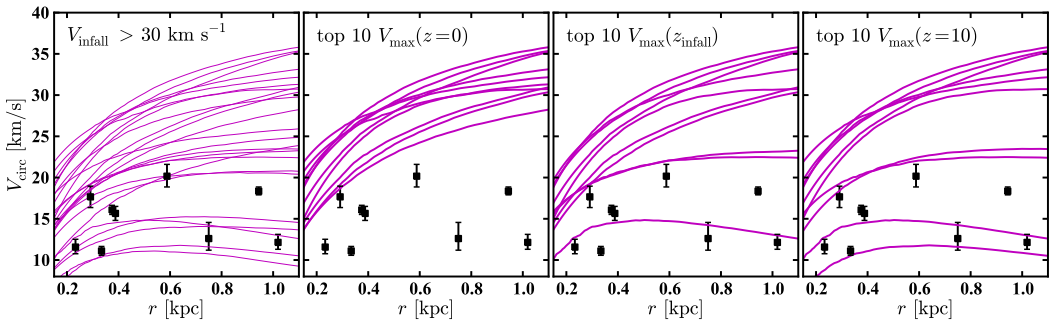

Figure 3: *First panel*: Circular velocity profiles for dSph host candidates ($V_{\rm infall} = V_{\rm max}(z_{\rm infall}) > 30$ km $\cdot$ s^{-1}, $V_{\rm max}(z = 0) > 10$ km $\cdot$ s^{-1}), excluding Magellanic Cloud candidates ($V_{\rm infall} > 60$ km $\cdot$ s^{-1}, $V_{\rm max}(z = 0) > 40$ km $\cdot$ s^{-1}), in the Aquarius E halo ($M_{\rm vir} = 1.4 \times 10^{12} M_\odot$). The black data points with error bars are the nine bright Milky Way dSphs (those with $L_V > 10^5 L_\odot$, excluding the Sagittarius dwarf, which is not currently in equilibrium). *Second panel*: The 10 dSph host candidates with the largest current maximum circular velocities, and therefore the largest present-day mass. *Third panel*: The 10 dSph host candidates with the largest maximum circular velocities at the time of infall onto the galactic halo, when the subhalo mass is maximized. *Fourth panel*: The 10 dSph host candidates with the largest maximum circular velocities at $z = 10$, close to the time of reionization. The most massive halos at all three times manifestly fail to be compatible with the halos that host the Milky Way's dSphs. Analogous results are found for the other five Aquarius halos. Reproduced from [6].

Figure 3 shows that the largest subhalos today, at the time of infall onto the galactic halo, and at the epoch of reionization, are dynamically incompatible with the observed Milky Way dSphs. Short of an extreme downward revision of the currently accepted mass of the Milky Way, to much less than $10^{12} M_{\odot}$, or extremely large and highly stochastic supernova feedback which preferentially removes baryons in the largest halos, it seems that some sort of new physics at small scales is required to account for this discrepancy [6].

2.3.3 The Core/Cusp Problem

Although the strongest evidence for dark matter at this point probably comes from the CMB, it is classically justified as a mechanism to explain observations of galaxy rotation curves, which show flat rather than decreasing circular velocity profiles within spiral galaxies, i.e. they predict $\alpha = -2$ at intermediate distances, where $\rho \propto r^{\alpha}$. However, ΛCDM simulations of the innermost parts of galaxies have for several decades consistently been in conflict with observations. Already in the early 1990s, the first simulation [86] which was able to resolve the inner regions of a DM halo suggested that $\alpha = -1$ there, while H_I observations of dwarf galaxies [7, 87] strongly preferred an cored isothermal profile, with constant central density, $\alpha = 0$. A systematic survey of CDM halos in various cosmologies by Navarro, Frenk, and White [88] showed all of them were well fit by a profile with an inner slope $\alpha = -1$ and an outer slope $\alpha = -3$, which has become known as an "NFW profile:"

$$\rho_{\mathrm{NFW}}(r) = \frac{\rho_s}{(r/r_s)\,(1 + r/r_s)^2}, \tag{27}$$

where r_s and ρ_s are the characteristic radius and density of the halo. The next generation of simulations [8, 89] favored even steeper profiles, $\alpha_{\mathrm{inner}} = -1.5$ and $\alpha_{\mathrm{outer}} = -3$.

More recent large-scale simulations have tended to find a gradual turn-over in slope at small radii, represented by an exponential "Einasto profile" [90, 91]:

$$\rho_{\mathrm{Ein}}(r) = \rho_{-2,\mathrm{Ein}} \exp\left(-2n\left[(r/r_{-2,\mathrm{Ein}})^{1/n} - 1\right]\right), \tag{28}$$

where r_{Ein} is the radius where the density profile has a logarithmic slope of -2, and $\rho_{-2,\mathrm{Ein}}$ is the density at this radius. For typical Einasto parameter values $4 < n < 8$, $r_{-2,\mathrm{Ein}} = 10$ kpc,

we find $\alpha = -1.3 \pm 0.2$ at 1 kpc, -0.9 ± 0.2 at 0.1 kpc. Two recent high-resolution simulations [92, 93] found good fits to Einasto profiles, with $\alpha \approx -0.8$ at 0.1 kpc. At the same time, more precise H_I data of seven dwarf galaxies from the THINGS survey [94] gives $\alpha = -0.29 \pm 0.07$, confirming that $\alpha \neq 1$ is a real effect and not due to systematic effects like pointing offsets, beam smearing, or non-circular orbits [95, 96, 97] which had previously prevented that conclusion.

Despite some convergence between pure-CDM simulations and observations, then, there is still a significant discrepancy between the two. Efforts to resolve the problem within the standard picture have focused on baryonic effects on the central density profile. Over extended periods of time, baryons in the interstellar medium will tend to dissipatively settle towards the center of the galaxy, pulling dark matter with it via so-called "adiabatic contraction" [98]. This will *steepen* the innermost parts of the density profile, worsening the discrepancy between simulation and observation. One suggestion for flattening inner profiles is very rapid mass loss due to starburst-triggered outflows [99, 100], although this is insufficient to account for all of the necessary flattening [101]. Another possibility is the disruption of cusps during formation by the dynamical friction of merging gas clouds, which can be highly efficient at early times [102]. Most recently, a simulation [103] incorporating a multi-phase interstellar medium, which allowed for detailed modeling of star formation and supernovae-driven outflows, found that the most important effects are the wind-driven loss of low-angular momentum gas from central regions [104], which inhibits bulge formation, and fluctuations in the gravitational potential due to random bulk motion of gas and rapid outflows [105, 106], which result in a shallower density profile. These effects apparently reduce the inner slope to $\alpha \sim -0.4 \pm 0.1$ [107], in good agreement with the THINGS experimental result.

3 Self-Interacting Dark Matter: Theory

3.1 The Basic Picture

The basic idea of Spergel and Steinhardt [10] was to consider dark matter with a non-negligible scattering cross section per unit mass[3] σ such that a typical SIDM particle at the solar radius, moving at $v_0 \approx 220$ km $\cdot$ s^{-1} in a region with a local DM density [52] $\rho_0 = 0.39 \pm 0.03$ GeV $\cdot$ cm^{-3} would undergo a non-negligible number of collisions in a Hubble time. Requiring at least 1 collision per Hubble time on average sets an upper bound on the local mean free path λ_0:

$$\lambda_0 \lesssim \frac{v_0}{H_0} = 2.2h^{-1} \text{ Mpc}. \tag{29}$$

At the same time, a lower bound is set by requiring that, as observed, clusters are triaxial rather than spherical on large scales, i.e. there is less than one collision per Hubble time at the half-mass radius of a DM halo. The half-mass radius is of comparable magnitude to the virial radius, at which the density is approximately 200 times the mean matter density of the universe [108], while the velocity is approximately the same as at the solar radius (the motivation for dark matter in the first place!), so

$$\lambda_0 \gtrsim \frac{v_0}{H_0} \cdot \frac{200\Omega_m \rho_{c,0}}{\rho_0} = 1.51h^{-1} \text{ kpc}, \tag{30}$$

using (2) $\Omega_m = \Omega_b + \Omega_c \approx 0.13h^{-2}$ and the value of $\rho_{c,0}$ from (9). The cross section per unit mass is related to the mean free path by

$$\lambda(r) = \frac{1}{\rho(r)\sigma}, \tag{31}$$

so the corresponding range of σ values that have an effect at the solar radius but not the half-mass radius is

$$0.21h \text{ cm}^2 \cdot \text{g}^{-1} \lesssim \sigma \lesssim 30.9h \text{ cm}^2 \cdot \text{g}^{-1}, \tag{32}$$

[3] Throughout the rest of this paper, we will use σ to denote a cross section *per unit mass*, as is the standard practice in the SIDM literature (e.g. [13, 23, 24]), rather than using it as a cross section as we did in section 2 above.

using $\rho_0 = 0.39\ \mathrm{GeV} \cdot \mathrm{cm}^{-3}$ [52].

Endowing the dark matter with a scattering cross section between these two extremal values should not affect observations on large scales, assuming that, as in the standard ΛCDM picture, it becomes nonrelativistic with the correct cosmological abundance[4] by the time of BBN and recombination. There should be no effect on the CMB or BAO even if the SIDM is optically thick when they are formed, since structure formation is still nonlinear and inhomogeneities are small so scattering has little overall effect. The requirement of negligible collisions per Hubble time at the virial radius (30) ensures that a nonzero σ will have no effect on the first stages of structure formation, e.g. on halo collapse and formation. In particular, this ensures that halos should remain triaxial on large scales. However, at smaller radii (and thus at correspondingly larger densities), the SIDM will become collisional: at radii where the dark matter becomes optically thick,

$$\tau(r) = \frac{r}{\lambda(r)} \gg 1, \tag{33}$$

there are many scatterings within an orbit, so that orbits become distributed isotropically rather than radially, increasing the phase space entropy and therefore leading to a shallower inner density profile. This has the potential to solve the core/cusp problem without recourse to baryonic feedback: the dark matter will naturally assume an isothermal profile, $\alpha \sim 0$.

Furthermore, the addition of a scattering cross section means that dark matter particles within subhalos can interact directly with particles within the main halo. Since subhalos have much smaller velocity dispersions, most two-body scatterings between satellite and halo particles will lead to both particles being ejected from the subhalos, gradually destroying substructure, as needed to resolve the missing satellite problem. Dwarf halos with the highest central densities will preferentially survive, since particles in optically thick areas will be prevented from escaping—in agreement with the observation that the Milky Way's bright dSphs have very high phase space densities [74].

[4]Of course, there is now no longer any reason why the WIMP miracle should apply, so it is unlikely that thermal SIDM production will lead to the correct relic abundance. SIDM candidates from particle physics therefore tend to have non-thermal production mechanisms, as discussed in Section 3.2 below.

3.2 Particle Physics Candidates

All viable dark matter candidates must have some production mechanism which leads to the correct abundance—production as a thermal relic, in the case of WIMPs. To resolve the core/cusp and missing satellite problems, SIDM candidates must also have a cross section per unit mass that lies in the viable range (32). This subsection presents some of the proposed SIDM candidates from particle physics which satisfy these requirements.

3.2.1 Mirror Matter

Already in the original paper [10], Spergel and Steinhardt pointed out that any SIDM particle making up the entirety of the dark matter must have a cross section satisfying (31)

$$\sigma m_\chi = \frac{m_\chi}{\rho_0 \lambda_0} = 8.31 \cdot 10^{-25} \text{ cm}^2 \left(\frac{m_\chi}{1 \text{ GeV}}\right) \left(\frac{\lambda}{1 \text{ Mpc}}\right)^{-1}, \tag{34}$$

which is intriguingly close to the typical hadron cross section, 1 b $= 10^{-24}$ cm^2, when $m_\chi \approx m_{\text{baryon}}$. Hence a stable "dark baryon" with an identical mass and cross section to the proton or neutron but only gravitational interactions with ordinary baryonic matter could act as a natural SIDM candidate. In "mirror world" theories [109, 110, 111], parity symmetry, experimentally observed to be broken in weak interactions [112, 113], is exactly restored by adding an additional mirror sector, which initially has particle content identical to the standard particles. Symmetry breaking can increase the mass of the mirror particles [114, 115], naturally explaining why the dark matter abundance is of the same order of magnitude but larger than the baryon abundance (2). In one such model of asymmetric symmetry breaking, mirror atomic hydrogen turns out to have the desired cross section [116], and is therefore a natural SIDM candidate. Alternatively, symmetry breaking could alter the couplings in the mirror sector such that the mirror neutron is stable rather than the mirror proton; it would then have the desired properties to be the SIDM [117].

3.2.2 Q-Balls

Another possibility [118] is that the SIDM is made up of "Q-balls" [119, 120, 121, 122], non-topological solitons (e.g. energy-minimizing semiclassical field configurations) in quantum field theories of scalar fields ϕ with a conserved global U(1) charge. Because they are not

point particles, Q-balls can evade the unitarity bound for s-wave scattering of point particles which follows from the optical theorem [14],

$$\sigma m_\chi \leq \frac{16\pi}{(m_\chi v_{\rm rel})^2}, \tag{35}$$

where $v_{\rm rel}$ is the relative velocity of the scattering SIDM particles. This gives a corresponding upper bound on m_χ: taking the lowest interesting value of σ in (32) and $v_{\rm rel} \sim 1000\,{\rm km}\cdot{\rm s}^{-1}$, as in clusters, gives

$$m_\chi \leq 18.9\ {\rm GeV}, \tag{36}$$

taking $h \sim 0.7$ [1]. However, because Q-balls are not pointlike there is a much wider range of allowed masses for particles with the necessary values of σ. In addition, it is possible that interacting Q-balls could stick together after colliding, ensuring that each particle undergoes only a few collisions per particle, which would significantly increase the allowed upper bound on σ in (32) [118]. Q-balls can be produced in sufficient abundance to make up the dark matter by a variety of mechanisms, such as via second-order phase transitions [123], "solitosynthesis" [124, 125], a process analogous to nucleosynthesis resulting from a universal Q charge asymmetry, or the "Affleck-Dine mechanism" [126, 127, 128], the fragmentation of a coherent condensate of scalar fields, e.g. from supersymmetry breaking during reheating at the end of an inflationary phase. Thus the Q-ball parameter space contains regions where they have the correct properties and abundance to make up the SIDM, but such areas are not naturally selected.

3.2.3 Singlet Scalar

The simplest renormalizable addition to the particle content of the Standard Model consistent with observations is a singlet scalar S [129], neutral under the SM gauge group, which adds three new parameters: an unrenormalized mass m_0 plus a dimensionless self-coupling λ_S and a dimensionless Higgs coupling λ_h. For generic values of its physical mass, m_S, smaller than a few TeV, S has the correct abundance at freezeout for $\lambda_h = \mathcal{O}\,(0.1\text{–}1)$ [130, 131]. In this sense, S is a natural DM candidate, without the need for tuning. The self-coupling parameter λ_S can then be freely adjusted, but translating large λ_S into a large cross section also requires a small renormalized S mass ($m_S \lesssim 1$ GeV) [132], which requires

part-per million tuning of m_0 vs. $\lambda_h v^2$ [133], where v is the Higgs VEV. Hence S requires extreme tuning to be a viable SIDM candidate.

3.2.4 Hidden Sectors

String theory models generically contain hidden sectors. If the hidden gauge group is non-Abelian, e.g. SU(N), then its analogues of glueballs could be semi-stable, with lifetimes much longer than the present age of the Universe, and interact very weakly with ordinary matter but strongly amongst themselves [134, 135], acting as SIDM with the right cross section if their masses are $\mathcal{O}(1)$ GeV. However, deriving the correct abundance is more difficult. To avoid spoiling nucleosynthesis, the exotic gluons must never have been in equilibrium with ordinary matter, so they must be produced non-thermally, e.g. by inflationary reheating to $\sim 10^{12\text{–}14}$ GeV, below the equilibrium temperature [135]. There exist examples of realistic string models, e.g. certain free fermionic models [136, 137], which contain the necessary SU(2) or SU(3) hidden gauge factors and allow the correct hidden gluon confinement scales to produce the correct DM abundance, although at the cost of substantial fine-tuning [135].

3.2.5 Inflaton Fields

In the braneworld scenario [138, 139, 140], where the observable universe is a manifold embedded in a space of higher dimension, the Friedmann equation has a quadratic density term, allowing steep inflationary potentials [141]. As a result, the universe can reheat via gravitational particle production rather than via inflaton decay [142], and the inflaton may remain stable or metastable instead of decaying. In this case the inflaton could be identified with the dark energy [143] or as a dark matter candidate [144]. If the inflaton oscillates around a post-inflationary minimum with sufficiently high amplitude, it can gain a mass and energy density that account for the observed dark matter abundance. For some choices of the braneworld inflationary potential, the dark matter could be endowed with the right scattering cross section for an SIDM candidate [145, 146]. However, the amount of gravitational particle production from reheating in models of this type spoils nucleosynthesis [147]. Reheating via production and decay of primordial black holes [148, 149] is an alternative, but a particular initial mass function must be assumed to derive the correct DM abundance [147].

3.2.6 Hidden Charged Dark Matter

In supersymmetric models that undergo gauge-mediated supersymmetry breaking [150, 151] in a single hidden sector, then the hidden sector mass scale m_X and gauge couplings g_X are related to the weak scale and coupling[5]:

$$\frac{m_X}{g_X^2} \sim \frac{m_w}{g_w^2}, \tag{37}$$

so

$$\Omega_X \sim \langle \sigma v \rangle^{-1} \sim \frac{m_X^2}{g_X^4} \sim \frac{m_w^2}{g_w^4} \sim \Omega_{DM}, \tag{38}$$

the "WIMPless miracle" [152, 153]. Hence stable hidden particles naturally give the correct relic density, even with hidden sector coupling constants far from g_w. In general, the hidden sector will not contain any stable particles, so the WIMPless miracle will not produce a dark matter candidate. If the hidden sector has an *exact* U(1) symmetry, however, the lightest charged hidden particle will be a natural DM candidate, so-called "hidden charged dark matter" [154]. Because the DM candidate is charged, it will undergo Rutherford scattering, with some choices of the hidden sector parameters giving a scattering cross section within the interesting SIDM range (32) [154].

3.3 Extensions of SIDM

The observational constraints presented in Section 4 below essentially exclude the astrophysically interesting parameter space (32) for SIDM with a constant cross section σ. Attempts to preserve the basic picture—i.e. to salvage SIDM as a resolution to the missing satellites and core/cusp problems—have therefore usually relied on proposing a more complicated functional form for the cross section as a function of velocity. In response to simulations (discussed in subsection 4.2.1 below) which suggested that cluster-scale SIDM halos developed flat cores incompatible with observations, Yoshida *et al.* [12] suggested introducing an energy-dependent cross section to reduce the effectiveness of high-energy scattering; they proposed

$$\sigma \propto \frac{1}{v}, \tag{39}$$

[5]Following [153], we write $\langle \sigma v \rangle \sim g^4/m^2$, rather than g^2/m^2 as we did in (21). This is simply a matter of definition, since g is a dimensionless quantity; the results are consistent.

which would make the collision rate at the scale radius independent of halo mass. Hui [14] pointed out that $\sigma \propto v^{-1}$ in the low-velocity limit required *inelastic* scattering. Hennawi and Ostriker [13] considered a general power-law scaling,

$$\sigma(v) = \sigma_0 \left(\frac{v}{v_0} \right)^{-a}, \tag{40}$$

and discussed limits in the σ–a plane. The remainder of this subsection presents a more detailed modification motivated by particle physics, before introducing the fractional SIDM which will be our concern for the rest of the paper.

3.3.1 Yukawa-Potential Interacting Dark Matter

Instead of the hard-sphere scattering typically considered in SIDM calculations, Loeb and Weiner [15] considered $\chi\chi$ scattering through exchange of a massive scalar or vector ϕ particle, i.e. via a Yukawa potential, hence the name "Yukawa-Potential Interacting Dark Matter" (YIDM). This "dark force" was previously considered [155] in the context of possible signatures of strong DM annihilation in high-energy cosmic rays [156, 157] and γ-ray observations of the galactic center [158, 159]. Exchange of a light ϕ could lead to a Sommerfeld enhancement [160] of the $\chi\chi$ annihilation cross section, which would necessarily increase the $\chi\chi$ scattering cross section as well [161, 162].

Elastic scattering through a massive mediator is equivalent to Coulomb scattering in a plasma [163], fit by a velocity-dependent cross section [15]:

$$\sigma m_\chi \approx \begin{cases} \frac{4\pi}{m_\phi^2} \beta^2 \ln\left(1 + \beta^{-1}\right) & \beta \lesssim 0.1, \\ \frac{8\pi}{m_\phi^2} \beta^2 / \left(1 + 1.5\beta^{1.65}\right) & 0.1 \lesssim \beta \lesssim 10^3, \\ \frac{\pi}{m_\phi^2} \left(\ln\beta + 1 - \frac{1}{2}\ln^{-1}\beta\right)^2 & \beta \gtrsim 10^3, \end{cases} \tag{41}$$

(valid in the classical limit, where the de Broglie wavelength is shorter than the ϕ Compton wavelength) where β is a dimensionless velocity,

$$\beta = \frac{\pi v_\sigma^2}{v^2} = \frac{y^2}{2\pi} \frac{m_\phi}{m_\chi} v^{-2}, \tag{42}$$

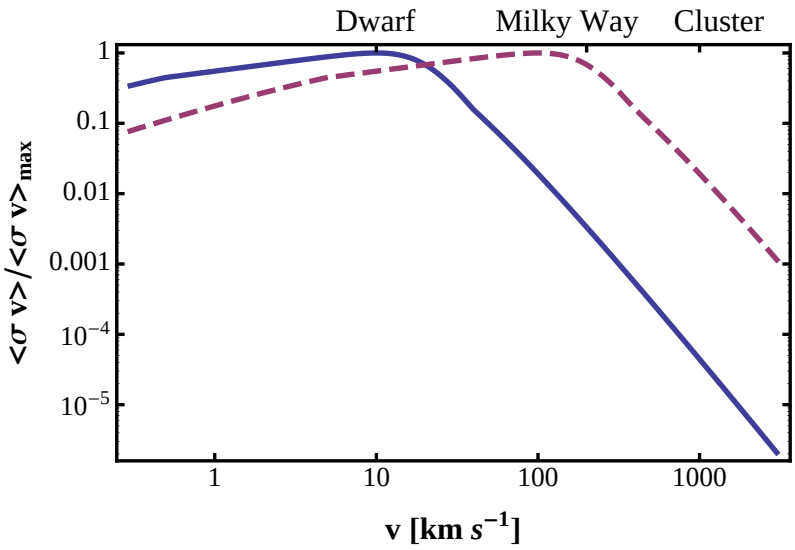

Figure 4: Normalized YIDM momentum-weighted scattering cross section plotted as a function of velocity. The blue solid curve peaks at typical dwarf galaxy circular velocities, $v_\sigma = 10\ \mathrm{km \cdot s^{-1}}$, while the purple dashed curve peaks at typical Milky Way circular velocities, $v_\sigma = 100\ \mathrm{km \cdot s^{-1}}$. Note the logarithmic scale: the effects of scattering rapidly become negligible for velocities larger than v_σ. Reproduced from [15].

with y the coefficient of the $\chi\bar{\chi}\phi$ interaction and v_σ the velocity at which the momentum-weighted scattering rate $\langle \sigma v m_\chi \rangle$ peaks, at a cross section of

$$\sigma_{max} m_\chi = \frac{8\pi}{m_\phi^2} \frac{\pi^2}{1 + 1.5\pi^{1.65}} \approx \frac{22.72}{m_\phi^2}. \tag{43}$$

Figure 4 shows the YIDM scattering rate as a function of velocity for several choices of v_σ, which can be chosen by tuning the particle physics parameters y, m_ϕ, m_χ via the relation (42). The functional form of the YIDM scattering cross section (41) ensures that the scattering rate falls off rapidly for velocities larger than v_σ, such that a choice of v_σ essentially sets the circular velocity scale (and therefore the mass/distance scale) at which the χ self-interactions become important. A recent resimulation of one of the Aquarius halos using YIDM with $v_\sigma = 10\ \mathrm{km \cdot s^{-1}}$ and $v_\sigma = 30\ \mathrm{km \cdot s^{-1}}$ [164] finds that the concentrations of the most massive subhalos in a Milky-Way sized galactic halo are significantly reduced, partially alleviating the mismatch between Milky Way dSphs and massive subhalos [6] shown in Figure 3 above.

Finally, it is possible that the YIDM scattering could be *inelastic*: that is, the χ particles could have excited states χ^*, χ^{**}, etc. Consider the simplest case, where only one excited state exists. Already, the presence of the excited state has introduced another scale into the problem: δ, the energy splitting between ground and excited states. In the general case, inelastic dark matter (iDM) [165, 166, 167] can have drastic effects for direct detection experiments, as well as potentially accounting for the observed deficit of dwarf galaxies [15].

3.3.2 Fractional SIDM

Subsection 2.3 above discussed the conflicts between simulations and observations that make the simplest assumption, that the dark matter is monolithic, cold, and collisionless, problematic. Given that this simple model of CDM exhibits difficulties, it is natural to consider more complicated arrangements. Naively, since the dark matter abundance is greater than the abundance of ordinary matter, we might well expect the dark sector to be at least as complex as the Standard Model content we observe.

As a first step towards confronting this potential complexity, in the remainder of this paper we assume that the dark matter is composed of two species. We take the vast majority of the dark matter to be the standard collisionless CDM, making no assumptions about its mass or nongravitational interactions. However, we assume that a small mass fraction f (which we will typically take to be $f \sim 0.01$) of the total dark matter is self-interacting, with a constant elastic scattering cross section per unit mass σ_0. Following the outline of the original SIDM proposal [10], we make the assumption that σ_0 is sufficiently small that the self-interaction does not play a role in the initial formation of galactic halos, so that the initial density profile of the SIDM is identical to the density profile of the dark matter as a while. (We will quantify this assumption in subsection 4.1.1 below.)

Because f is small, observations on the scale of clusters, galaxies, or galactic cores will not reveal any sign of the fractional SIDM's presence; in the next section, we review the existing constraints on SIDM and demonstrate that most of them become unimportant when $f \ll 1$. A corollary of this insensitivity to substructure observations is that fractional SIDM (at least with small f, as we consider here) cannot resolve the ΛCDM problems considered in subsection 2.3 above. However, fractional SIDM may be useful in resolving another difficulty in the standard picture: the existence of supermassive black holes at large

redshifts. Accordingly, starting in section 5 and for the rest of the paper, we turn our attention to black hole formation, evolution, and related matters.

4 Self-Interacting Dark Matter: Constraints

In this section, we review the existing observational and theoretical constraints on self-interacting dark matter, re-evaluating them where newer data has become available[6]. In addition, we generalize the constraints to the fractional case, when the SIDM makes up a mass fraction $f < 1$ of the total dark matter in the system, and show that most of them are either significantly weakened or disappear entirely.

4.1 Galaxy Constraints

4.1.1 Galaxy Formation Constraint

Following [13], we assume that the initial collapse of collisionless dark matter will result in halos with a generalized NFW density profile, also known as a Zhao profile [168]:

$$\rho(r) = \frac{\rho_s}{(r/r_s)^{\alpha}(1 + r/r_s)^{\epsilon-\alpha}}, \tag{44}$$

where ρ_s and r_s are the characteristic density and scale radius, respectively, and α and ϵ parameterize the inner and outer slopes. Following [13] and the results for fits to NFW-like profiles [8, 89] reported in subsection 2.3.3, we take $\alpha \sim 1.3 \pm 0.2$. This will be a good approximation even if (some of) the dark matter is collisional when the *dynamical* timescale of collapse is much less than the *relaxation* timescale due to collisions at the characteristic radius:

$$t_{\rm dyn}(r_s) \ll t_{\rm rel}(r_s) \approx \frac{1}{\tau_s} t_{\rm dyn}(r_s) \rightarrow \tau_s \ll 1; \tag{45}$$

i.e. so long as the halo is optically thin at its characteristic radius. Where $\tau \ll 1$, SIDM particles have not had the chance to undergo any self-interaction, so we are justified in assuming they follow the same profile, $\rho_{\rm SIDM}(r) = f\rho(r)$.

[6] Some of the constraints, particularly those in [13], were formulated in terms of a velocity-dependent cross section, $\sigma \propto v^{-a}$. Because this paper is primarily concerned with constant-cross section SIDM, we set $a = 0$ when reporting results.

M_Δ ($M_\odot$)	c_Δ	ρ_s (g/cm^3)	r_s (kpc)	σ_0 (cm^2/g) (upper limit)
10^8	13.14	$4.00 \cdot 10^{-25}$	0.92	$878.5 f^{-1}$
10^9	12.23	$3.34 \cdot 10^{-25}$	2.14	$453.9 f^{-1}$
10^{10}	11.26	$2.72 \cdot 10^{-25}$	5.00	$238.2 f^{-1}$
10^{11}	10.15	$2.11 \cdot 10^{-25}$	11.95	$128.6 f^{-1}$
10^{12}	8.91	$1.54 \cdot 10^{-25}$	29.32	$71.9 f^{-1}$
10^{13}	7.52	$1.02 \cdot 10^{-25}$	74.82	$42.3 f^{-1}$
10^{14}	5.96	$5.93 \cdot 10^{-26}$	203.33	$26.9 f^{-1}$
10^{15}	4.25	$2.75 \cdot 10^{-26}$	614.05	$19.2 f^{-1}$

Table 1: Density contrasts, NFW profile characteristic densities and radii, and maximum allowed cross section for NFW-like halo formation for a range of halo masses.

In general $\tau = \rho r \sigma$, so clearly in the fractional case $\tau_s = f \rho_s r_s \sigma_0$. This gives our first constraint,

$$\sigma_0 \leq \frac{1}{f \rho_s r_s}. \tag{46}$$

Note that this is *not* an absolute constraint on σ_0: it just represents the maximum allowed value for which the assumption of an initial NFW density profile at the time of formation is valid.

In the formalism of [13], ρ_s and r_s are uniquely determined by the halo virial mass M_Δ and concentration c_Δ:

$$\rho_s = \delta_c \rho_{\rm crit} = \delta_c \frac{3H^2}{8\pi G}, \tag{47}$$

with δ_c given by

$$\delta_c = \frac{\Delta}{3} \frac{c_\Delta^3}{\ln(1+c_\Delta) - c_\Delta/(1+c_\Delta)}, \tag{48}$$

where $\Delta \approx 178 \Omega^{0.45}$ for a flat universe, with Ω the matter density, and

$$r_s = \frac{r_\Delta}{c_\Delta} \approx \frac{9.52 \times 10^{-2}}{c_\Delta} \left(\frac{M_\Delta}{M_\odot} \right)^{1/3} h^{-2/3} \Delta^{-1/3} \text{ kpc}, \tag{49}$$

where r_Δ is the virial radius and $H = 100h$ km/s/Mpc. Eke, Navarro, and Steinmetz [169] performed simulations to study the dependence of c_Δ on the halo mass M_Δ and the underlying cosmology: we use their publicly available code [170] and the WMAP7 cosmological parameters [1] to evaluate $c_\Delta(M_\Delta)$ and thus determine the upper bound on σ_0 as a function of M_Δ and f.

Table 1 shows the computed value of c_Δ and resulting limit on σ_0 for a range of halo masses. The upper limit is a decreasing function of halo mass, but one that decreases very slowly: by only a factor of 50 over seven decades of mass. In subsection 6.4 below, we will model gravothermal collapse by assuming NFW initial conditions; the values in the table determine the range in which this assumption is valid.

4.1.2 Black Hole Accretion Constraint

Self-interacting dark matter has a non-negligible scattering cross section, which means that individual particles can lose energy and accrete onto seed black holes in the galactic center. Qualitatively, it is clear that a larger scattering cross section leads to larger black holes; this means that we can place a constraint on σ by observing the size of central black holes relative to their host galaxies.

To make this constraint quantitative, assume that the halo is in hydrostatic equilibrium[7]. In the inner region ($r \ll r_s$), the density goes as

$$\rho(r) \approx \rho_s (r/r_s)^{-\alpha}, \tag{50}$$

so [13] the velocity dispersion is

$$v^2(r) \approx v_s^2 \left(\frac{r}{r_s}\right)^{2-\alpha}, \tag{51}$$

with

$$v_s^2 = \mu G \rho_s r_s^2 \tag{52}$$

$$\mu \equiv \frac{2\pi}{(3-\alpha)(\alpha-1)}. \tag{53}$$

Now assume there exists an accreting black hole at early times ($t \ll t_{\rm rel}$), with Bondi accretion radius [171]

$$r_c = \frac{G M_{\rm BH}}{v^2(r_c)}. \tag{54}$$

[7] In the case of multi-component dark matter, we assume that each individual component is separately in hydrostatic equilibrium, as discussed in subsection 6.3 below.

Inside r_c, the density and velocity dispersion will go as [13]

$$\rho(r) \approx \rho(r_c) \left(\frac{r}{r_c}\right)^{-3/2}, \tag{55}$$

$$v^2(r) \approx v^2(r_c) \left(\frac{r}{r_c}\right)^{-1}. \tag{56}$$

For simplicity, first consider $f = 1$. Then the optical depth is

$$\tau(r) = r\rho(r)\sigma_0 \approx \begin{cases} \tau_c \left(\frac{r}{r_c}\right)^{(\alpha-1)/2}, & r \lesssim r_c, \\ \tau_s \left(\frac{r}{r_s}\right)^{1-\alpha}, & r_c \lesssim r \ll r_s, \end{cases} \tag{57}$$

where $\tau_c \equiv \rho(r_c) r_c \sigma_0$, $\tau_s \equiv \rho_s r_s \sigma_0$, and we have used equations (50) and (55) and assumed $r_c \ll r_s$. Growth via Bondi accretion will only occur when the SIDM can be treated as an adiabatic gas, $\tau \gg 1$. We place a conservative lower bound on the black hole mass by assuming that all of its mass comes from Bondi accretion, ignoring a smaller contribution from later diffusion. In this case, the black hole will grow until $\tau(r_c) \approx 1$, so that

$$M_{\rm BH} = \frac{v^2(r_c) r_c}{G} \approx \mu \rho_s r_s^3 \left(\frac{r_c}{r_s}\right)^{3-\alpha} \geq \mu \rho_s r_s^3 \tau_s^{(3-\alpha)/(\alpha-1)}, \tag{58}$$

where in the last step we have used the fact that $\tau(r_c) \approx 1$. Solving (58) for σ_0 gives the desired constraint,

$$\sigma_0 \leq \left(\frac{M_{\rm BH}}{\mu \rho_s r_s^3}\right)^{(\alpha-1)/(3-\alpha)} \frac{1}{\rho_s r_s}. \tag{59}$$

In the case that only a portion of the dark matter is self-interacting, we must take $\rho_s \to f\rho_s$, so the inequality becomes

$$\sigma_0 \leq \left(\frac{M_{\rm BH}}{\mu \rho_s r_s^3}\right)^{(\alpha-1)/(3-\alpha)} \frac{f^{-2/(3-\alpha)}}{\rho_s r_s}. \tag{60}$$

Unsurprisingly, the most stringent constraints come from galactic halos with large radii (and thus large masses) but light central black holes. Hennawi and Ostriker [13] cite the example of M33, which has a black hole mass constrained to be $< 1500 M_\odot$ [172] and a halo mass

measured to be approximately $2.2 \times 10^{11} M_\odot$ [173]. Using these values gives

$$\sigma_0 \leq 37.79 f^{-1.05},\ 4.07 f^{-1.18},\ 0.26 f^{-1.33}\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1} \tag{61}$$

for $\alpha = 1.1,\ 1.3,\ 1.5$ respectively. This constraint will be relevant in later sections, where we will focus precisely on growing SMBHs via Bondi accretion. However, the contraint assumes the presence of an accreting black hole before core collapse occurs, while we will generate the seed black hole by core collapse itself. In addition, the values in (61) are determined by demanding that an atypically small black hole still conform to this constraint—an argument which is vulnerable to alternative environmental explanations, e.g. that core collapse was disrupted by a serendipitously timed merger, or even that the original host black hole was kicked out the galaxy, as discussed in subsection 5.3 below.

4.1.3 Core Collapse Constraint

On long enough timescales (i.e. after many relaxation times), SIDM halos will undergo core collapse once the flow of heat to the halo center halts. This is a problem if we want to use SIDM to explain cored galaxy profiles, since SIDM cores will only persist for a finite time before collapsing. Because no such collapsed halos have been observed, either $f \ll 1$, so core collapse is inherently unobservable, or all observed halos satisfy

$$C_1 t_{\mathrm{rel}}(r_s) = \frac{C_1 t_{\mathrm{dyn}}(r_s)}{\tau_s} \geq t_H - t_f, \tag{62}$$

where t_H and t_f are the Hubble time and time of halo formation, respectively, $C_1 \approx 4.76$ from simulations [174], and the dynamical time is given by $t_{\mathrm{dyn}}(r) = r/v(r)$. Using (52) gives

$$t_{dyn}(r_s) = \frac{1}{\sqrt{\mu G \rho_s}}, \tag{63}$$

independent of the relative SIDM density (assuming an initial NFW profile), while $\tau_s = f \rho_s r_s \sigma_0$, so

$$\sigma_0 \leq f^{-1} \frac{C_1}{r_s \rho_s^{3/2} \sqrt{\mu G}} \frac{1}{t_H - t_f}. \tag{64}$$

Using the analytic fitting formulae in [175], Hennawi and Ostriker [13] find that $t_f/t_H = 0.26,\ 0.33,\ 0.45$ for halos of mass $10^8 M_\odot,\ 10^{10} M_\odot,$ and $10^{12} M_\odot$, respectively, giving the constraints

$$\sigma_0 \leq \frac{24.64}{f},\ \frac{8.94}{f},\ \frac{4.38}{f}\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}. \tag{65}$$

Again, this is only a constraint on SIDM as a solution to the cuspy halo problem: in other uses of SIDM, e.g. as a source of primordial black holes via gravothermal collapse [17], we want the collapse time to be as fast as possible.

Note also that this result implies that $t_{\mathrm{rel}}/(t_H - t_f)$ varies by only a factor of five over four orders of magnitude in mass, and the trend continues even to $10^{15} M_\odot$: the assumption of a constant SIDM cross section implies that clusters and dwarf galaxies are just as relaxed [13].

4.2 Cluster Constraints

4.2.1 Cluster Core Constraint

Yoshida *et al.* [12] performed N-body simulations of isotropic hard-sphere scattering by constant-cross section SIDM in cluster-sized halos ($M_\Delta \approx 7.4 \cdot \times 10^{14} M_\odot$). Starting with the same initial conditions, they performed four simulations, one with non-interacting CDM and the others with cross sections of $\sigma = 0.1,\ 1.0,\ 10.0\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, respectively. Figure 5 shows the resulting mass distributions of the simulated clusters. At the smallest cross section used, $\sigma = 0.1\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, the cluster develops a $40h^{-1}$ kpc core.

Strong lensing observations [176] of CL 0024+1654 ($M \approx 1.7 \times 10^{14} M_\odot$) reveal a small, dense core with a $35h^{-1}$ kpc radius, while X-ray observations [177] of EMSS 1358+6245 ($M \approx 4 \times 10^{14} M_\odot$) place a limit $r_c < 37h^{-1}$ kpc (90% CL); evidently if the dark matter is entirely self-interacting its cross section must not be much larger than $0.1\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$.

Yoshida *et al.* found that cores formed only where the collision rate per particle at the cluster center exceeded one or two per Hubble time, in accordance with the heuristic arguments in subsection 3.1. Hence introducing fractional SIDM with $f \ll 1$ will form much smaller cores, in the regions where the rate per particle exceeded $(1\text{–}2)f^{-1}$ in the simulations of Yoshida *et al.* At $40h^{-1}$ Mpc scales, the mean collision count was ~ 1 for $\sigma = 0.1\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, ~ 8 for $\sigma = 1.0\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, and ~ 25 for $\sigma = 10\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$ [12].

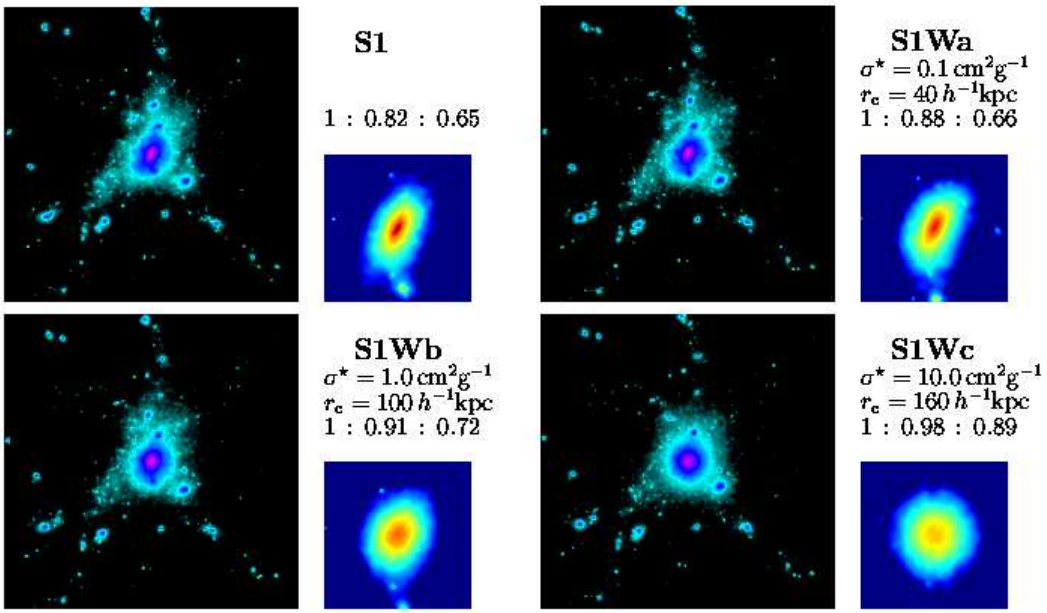

Figure 5: Properties of a simulated cluster ($M_\Delta \approx 7.4 \times 10^{14} h^{-1} M_\odot$) with varying SIDM cross sections per unit mass. For each value of σ (shown on the figure as σ^*), the core radius r_c and ratios of the three axes are reported. The large images show the mass distribution in a box with side length $15h^{-1}$ Mpc. The small figures show the central region ($2h^{-1}$ Mpc on a side), in a different color scale. Even the smallest value of σ leads to a $40h^{-1}$ Mpc core. Reproduced from [12].

Hence the experimental bounds on core radii are comfortably evaded for $f = 0.01$, even for $\sigma > 10\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$.

4.2.2 Merger Constraints

Observations of the Bullet cluster (1E 0657-56) reveal an offset between the gas "bullet" and the dark matter centroid of the merging subcluster, significant to at least 2σ [178]. Based on X-ray observations, Markevitch *et al.* [11] assume that the subcluster has already passed through the main cluster at least once, so that the separation is due to stripping and deceleration of the gas ; under this assumption they use the observed offset to place a constraint on the dark matter self-interaction strength, since sufficiently strong interactions should have slowed the SIDM as it has the gas bullet. Of course, this constraint is only meaningful if $f \sim 1$: the observational error is sufficiently large that a small portion of the dark matter that had, indeed, been slowed down could not have been detected. Assuming that the main cluster and subcluster have the same gas fraction, the observed offset requires that the SIDM optical depth must be ≤ 1. The measured dark matter surface density is $\Sigma_s \approx 0.2\ \mathrm{g} \cdot \mathrm{cm}^{-2}$, so

$$\sigma_0 \leq \frac{1}{f\Sigma_s} \approx \frac{5}{f}\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}. \tag{66}$$

A second constraint comes from the observed subcluster velocity, $v_s = 4500^{+1100}_{-800}\ \mathrm{km} \cdot \mathrm{s}^{-1}$, which is consistent with the expected free-fall velocity onto the main cluster of $4400\ \mathrm{km} \cdot \mathrm{s}^{-1}$ at core passage, decelerating to $3500\ \mathrm{km} \cdot \mathrm{s}^{-1}$ at the current location [11]. Agreement suggests that the subcluster is experiencing negligible drag forces, which would result if SIDM collisions were frequent enough; this places a second constraint on σ_0. Markevitch *et al.* find that the net momentum loss per particle collision is

$$p_{\mathrm{both}} = mv_s \left[1 - \cos\alpha \left(\cos^2\alpha - \frac{V^2}{v_s^2}\sin^2\alpha_1\right)^{1/2} - \sin\alpha \left(\sin^2\alpha - \frac{V^2}{v_s^2}\cos^2\alpha_1\right)^{1/2}\right] \tag{67}$$

if both particles escape and

$$p_{\mathrm{one}} = mv_s \left[1 - \cos\alpha \left(\cos^2\alpha - \frac{V^2}{v_s^2}\sin^2\alpha_1\right)^{1/2}\right] \tag{68}$$

if only one particle escapes, where α is the scattering angle and $V \approx 1900\ \mathrm{km \cdot s^{-1}}$ is the critical velocity to escape the subcluster fiducial radius. In the CM frame the scattering is isotropic when quantum effects are unimportant, i.e.

$$\frac{mv_0 r}{2\hbar} \ll 1, \tag{69}$$

where r is the linear scale of the interaction and $v_0 = \sqrt{v_s^2 + V^2} \approx 4900\ \mathrm{km \cdot s^{-1}}$ is the collision velocity. In the case of hard-sphere scattering, where $\sigma = 4\pi r^2$, this requires

$$\left(\frac{\sigma_0}{1\ \mathrm{cm^2 \cdot g^{-1}}}\right)^{1/2} \left(\frac{mc^2}{1\ \mathrm{GeV}}\right)^{3/2} \frac{v_0}{4900\ \mathrm{km \cdot s^{-1}}} \ll \frac{4\sqrt{\pi}\hbar}{4900\ \mathrm{km \cdot s^{-1}}} \approx 6.41, \tag{70}$$

which is only valid if $m \ll 3.46\ \mathrm{GeV}/c^2$ if $\sigma_0 \approx 1\ \mathrm{cm^2 \cdot g^{-1}}$ or $m \ll 939\ \mathrm{MeV}/c^2$ if $\sigma_0 \approx 50\ \mathrm{cm^2 \cdot g^{-1}}$, i.e. it is invalid for typical weak-scale masses. However, non-isotropic scattering is unlikely to lead to observable effects such as beaming (or, equivalently, to a failure to detect SIDM despite its presence) absent an already anisotropic system.

In the isotropic case the average momentum lost by the subcluster in each collision is

$$\bar{p} = mv_s \left\{1 - 4\int_{\sin\alpha_e}^{1} x^2 \left[x^2 - \frac{V^2}{v_s^2}(1-x^2)\right]^{1/2} dx\right\} \approx 0.096 mv_s, \tag{71}$$

where $\alpha_e \approx 23^\circ$ is given by the requirement that both particles escape in collisions where $\alpha_e < \alpha < \pi/2 - \alpha_e$. Then the difference from free fall is given by

$$\Delta v = \frac{\bar{p}}{m} f \Sigma_m \sigma_0. \tag{72}$$

Requiring that the difference be less than $1000\ \mathrm{km \cdot s^{-1}}$ gives

$$\sigma_0 \lesssim \frac{\Delta v}{.096 v_s f \Sigma_m} \approx \frac{7.72}{f}\ \mathrm{cm^2 \cdot g^{-1}}, \tag{73}$$

assuming that the subcluster passed through the center of the main cluster, so that the main cluster surface density is $\Sigma_m \approx 0.3\ \mathrm{g \cdot cm^{-2}}$. Again, if $f \ll 1$ the momentum loss would not be detected, so this only provides a constraint when $f \sim 1$.

Finally, a third constraint comes from the requirement that the subcluster has not yet

evaporated due to SIDM collisions. The mass-to-light ratio of the subcluster [178] is a factor of 1.1 ± 0.3 from the equivalent ratio for the main cluster, so the subcluster cannot have lost more than $f_{\rm evap} \approx 0.3$ of its initial mass. Net loss of SIDM particles occurs if both particles in a two-body collision have subsequent velocities larger than the escape velocity $v_{\rm esc} \approx 1200$ km $\cdot$ s^{-1}, which occurs at scattering angles

$$\frac{v_{\rm esc}}{v_0} < \sin\frac{\theta}{2} < \sqrt{1 - \frac{v_{\rm esc}^2}{v_0^2}}. \tag{74}$$

Such angles make up a fraction of the total solid angle

$$\chi = 1 - 2\frac{v_{esc}^2}{v_0^2}, \tag{75}$$

which is also the probability of particle loss per collision, again assuming isotropic scattering. In one transit through the cluster center, the collision probability per particle is

$$\tau_m = \sigma_0 f \Sigma_m, \tag{76}$$

so that the fraction of particles lost is

$$f_{\rm evap} = \chi\tau_m = \sigma_0 f \Sigma_m \left[1 - 2\left(\frac{v'_{esc}}{v_0}\right)^2\right], \tag{77}$$

where $v'_{esc} \approx v_{esc}\left[1 + (1 + f_{\rm evap})^{1/2}\right]/2$ takes into account the possible mass decline during the course of the passage. Requiring that $f_{\rm evap} < 0.3$ gives

$$\sigma_0 < \frac{0.3}{f\Sigma_m}\left[1 - 2\left(\frac{v'_{esc}}{v_0}\right)^2\right]^{-1} \approx \frac{1.16}{f}\ {\rm cm}^2 \cdot {\rm g}^{-1}. \tag{78}$$

This is the most stringent constraint thus far, but once again a small SIDM mass fraction would not observably effect the mass-to-light ratio, making the constraint unimportant for $f \ll 1$.

Randall *et al.* [179] compared observations to numerical simulations of systems similar to 1E 0657-56 to set additional constraints (with $f = 1$). The lack of a measured offset

between the dark matter and galaxy centroids gives a limit $\sigma_0 \leq 1.25\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, while an improved measurement of the ratio of mass-to-light ratios, 0.84 ± 0.07, gives a slightly tighter limit of $\sigma_0 \leq 0.6\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, again assuming the initial mass-to-light ratios were equal. Assuming a non-zero impact parameter between the subcluster and the main cluster core slightly weakens the limits.

4.2.3 Evaporation Constraint

Just as galactic-scale halos undergo core collapse as a result of heat transfer from self-interaction, efficient heat transfer within a cluster could evaporate galactic subhalos. Gnedin and Ostriker [180] claim that such destruction of substructure should occur after one relaxation time at the location of the galactic subhalo. The limit on σ_0 will again be of the form (64), though now we (conservatively) set $C_1 \approx 2$. Taking a cluster like EMSS 1358+6245 [181] with a mass $M_\Delta \approx 4 \times 10^{14} M_\odot$ gives a stringent constraint:

$$\sigma_0 \leq \frac{1.97}{f}\ \mathrm{cm}^2 \cdot \mathrm{g}^{-1}. \tag{79}$$

This appears to be problematic for the fractional SIDM case, but note that only the SIDM portions of the galactic subhalos would be evaporated: the remainder of the CDM would persist, and substructure would remain. In short, for small enough f there is no way by which this evaporation could actually be detected observationally.

5 Supermassive Black Holes

5.1 The Observational Picture

Observations of stellar dynamics have established the presence of a massive ($\sim 4.4 \cdot 10^6 M_\odot$) black hole in the center of the Milky Way [182, 183], as well as in other nearby galaxies [184, 185]; it is now routinely assumed that they reside in the center of essentially all galaxies [186]. At large distances, where the central kinematics of galaxies cannot be resolved, black holes cannot be directly observed. However, there is strong evidence that quasars are powered by radiation from actively accreting supermassive black holes [187, 188, 189]. If the black hole powering the quasar is assumed to be radiating at the Eddington luminosity,

where the gravitational force on the accreting gas is balanced by the radiation pressure, than its mass can be determined via the relation

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} \mu_e m_p c}{\sigma_T}, \tag{80}$$

where μ_e is the mean atomic weight per electron of the accreting material and σ_T is the Thompson cross section,

$$\sigma_T = \frac{8\pi}{3}\left(\frac{e^2}{4\pi\epsilon_0 m_e c^2}\right)^2. \tag{81}$$

Hence a black hole's mass can be measured by observing its luminosity in a given spectral band and applying a bolometric correction, observationally deterimined by measuring the spectral energy distributions of nearby quasars [190, 191], to obtain the total luminosity. Luminosities $\gtrsim 10^{47}$ erg $\cdot$ s^{-1} correspond to black holes with mass a few $\times\, 10^9 M_\odot$, the most massive observed. To date, approximately two dozen quasars have been discovered at redshifts $z \gtrsim 6$ by the Sloan Digital Sky Survey [18, 19], as well as deeper small-area surveys like the Canada-France High-z Quasar Survey [20] and the UKIDSS Large Area Survey [22], corresponding to a number density of ~ 1 Gpc^{-3} [192].

If black holes mostly grow via gas accretion, then their growth rate is limited to the amount of accretion corresponding to the Eddington luminosity (80), $\dot{M}_{\rm Edd} = L_{\rm Edd}/c^2$. Hence continuous Eddington-limited growth is exponential with an e-folding rate given by the Salpeter time [188],

$$t_{\rm Sal} = \frac{\epsilon_r \sigma_T c}{4\pi G m_p} \approx \left(\frac{\epsilon_r}{0.1}\right) 45.1 \text{ Myr}, \tag{82}$$

assuming mainly accretion of hydrogen gas. Here ϵ_r is the radiative efficiency, the ratio of radiated energy to accreted mass, usually taken to be $\epsilon_r = 0.1$, the value calculated [193] for accretion of matter from a binary companion onto a Schwarzchild black hole.

If we assume that indeed $\epsilon_r = 0.1$, then we find that a black hole can have grown by a factor of $10^{7.5}$ in the entire age of the universe[8] up to $z \sim 7$, or a factor $10^{8.3}$ by $z \sim 6.5$. Hence plausible explanations for the presence of multi-billion solar mass black holes at high redshifts in the standard picture must have two components: a means of forming seed black holes with masses of at least $10^2 M_\odot$ soon after structure formation begins and an

[8] Realistically, of course, baryons do not begin to efficiently virialize in rare dark matter halos until at least $z \sim 50$, so the total growth factor is less by a factor of at least $10^{0.5}$.

explanation for how continuous near-Eddington limit accretion for ~ 800 Myr is possible. In the next two subsections, we review each of these aspects in turn.

5.2 Black Hole Seeds

5.2.1 Population III Remnants

After recombination, $z_* = 1091.36 \pm 0.91$, the universe remained neutral until reionization at $z = 10.6 \pm 1.2$ [1], when it was probably reionized by low-metallicity Population III stars [194]. Before reionization occured, the lack of a sizable cosmic UV background meant that H_2 molecules were not photodissociated, so could act as a coolant and prevent gas fragmentation in the absence of metals [195]. Hence the first generation of stars could have been significantly more massive than later generations; simulations which show fragmentation does not occur find $M_* > 100 M_\odot$ [196, 197, 198]. However, some recent simulations show that fragmentation does occur, resulting in the formation of binary star systems where the larger member of the binary has a mass $M_* \approx 50 M_\odot$ [199, 200], or that radiative feedback from the protostar removes mass from the accreting gas disk via photodissociation, halting accretion at relatively low masses ($M_* \lesssim 140 M_\odot$) [201]. Even if these hazards are avoided, in order to be useful for producing SMBHs it is important that collapse occur *rapidly*, i.e. that the angular momentum in the circumstellar disc be quickly shed.

In the low-metalllicity limit, there are two mass regimes which result in black hole remnants: $25 M_\odot \leq M_* \leq 140 M_\odot$ and $M_* \geq 260 M_\odot$ [202]. In the lower range, less than half of the stellar mass collapses into a black hole: a $100 M_\odot$ star leaves behind a $\lesssim 40 M_\odot$ black hole [203]. These black holes are probably too light to occupy a dynamically stable position, so may not settle in the center of their host halo [204]. In the intermediate regime, stars explode as pair-instability supernovae after ~ 3 Myr when they have exhausted their supply of helium, leaving no remnant black hole behind [205]. In the most massive stars, however, the core is sufficiently massive that explosive burning does not suffice to reverse the intial implosion due to pair instability [206], and all mass left behind after helium burning is incorporated into the resulting black hole.

5.2.2 Monolithic Collapse

Another possibility is to form $10^{5-6} M_{\odot}$ black holes directly [207, 208, 209]—that is, by accreting large amounts of gas onto a central black hole at super-Eddington rates. This requires the gas halo to contract adiabatically, rather than fragmenting and forming protostars. One means of avoiding this fragmentation is for the gas to remain hot, $T_{\rm gas} \geq 10^4$ K, which is only possible if molecular hydrogen is prevented from forming. As mentioned above, the pre-reionization UV background was typically not large enough to destroy H_2 via photodissociation, but rare gas halos located very close to a luminous neighbor might be exposed to the necessary background [210]. Alternatively, sufficiently optically thick hot gas could trap Lyman α photons, which could cause the atomic hydrogen equation of state to stiffen and inhibit fragmentation [211]. Finally, fragmentation is suppressed in highly turbulent systems, for example in highly metal-enriched gas at later epochs [212].

Even if fragmentation is prevented, however, in general conservation of angular momentum will lead to the formation of a rotationally supported disc, preventing super-Eddington accretion [213, 214]. Even scenarios which form black holes from extremely low angular momentum material require some amount of angular momentum transport to enable collapse [215]. Ultimately, disrupting the disc requires a dynamical instability, either global, such as repeated outward angular momentum transport by bar formation, the "bars-within-bars" mechanism [216], or local, in which the disk becomes self-gravitating and marginally stable, developing spiral structures which transport angular momentum outwards and cause central mass inflows [217, 218].

5.2.3 Runaway Stellar Dynamics

Finally, $\sim 10^{2-4} M_{\odot}$ black holes could form from stellar-dynamical rather than gas-dynamical processes [219, 220]. Self-gravitating clusters of approximately solar-mass stars are the prototypical system that undergoes gravothermal collapse [16]. In systems with uniform-mass stars, binary formation has generally been thought to act as an energy sink and disrupt the core collapse process [221, 222] in all but the most massive clusters, $N \gtrsim 10^{6-7}$ [223, 224], which are very rare at high redshifts [192]. However, realistic clusters with multiple species of stars with different masses are vulnerable to a mass segregation instability, where dynamical

friction causes the more massive stars to sink towards the center of a cluster and undergo collapse on a shorter timescale [225, 226]. In this case runaway growth can occur, leading to core collapse into a massive star on a timescale [227]

$$t_{cc} \approx 3 \text{ Myr} \left(\frac{R_h}{1 \text{ pc}} \right)^{3/2} \left(\frac{M_{cl}}{5 \times 10^5 M_\odot} \right)^{1/2} \left(\frac{10 M_\odot}{\langle m \rangle} \right), \tag{83}$$

where R_h is the initial half-mass radius of the cluster, M_{cl} is the total cluster mass, and $\langle m \rangle$ is the mean mass of a star in the cluster. Recent Monte Carlo simulations find that a central star is formed with a mass $\sim 10^{-3} M_{cl}$ [228], and can then produce an intermediate-mass black hole remnant as described in subsection 5.2.1 above.

5.3 SMBH Evolution

Once a seed black hole has been successfully formed, it must grow fast enough to reach the $10^9 M_\odot$ range by $z = 6$–7. In the standard picture, black holes grow via two main mechanisms, direct accretion of gas and merger with other black holes. We consider each mechanism in turn.

If the growth comes primarily from gas, then, barring periods of super-Eddington growth, as shown in subsection 5.1 above accretion must be continuously at or near the Eddington limit to reach the required mass range. Hence the black hole must continuously be provided with infalling gas as fuel. However, if the black hole is formed as a remnant of a supermassive star, intense UV radiation, possibly accompanied by supernovae feedback if the star is in the appropriate mass range, may photoionize and evacuate the halo gas surrounding the star. Recent simulations suggest that this is indeed the case [229, 230, 231], and that as a result accretion may be delayed by up to 100 Myr, sacrificing nearly an order of magnitude of potential Eddington-limited growth (82).

In addition, the Salpeter e-folding time (82) for black hole growth via gas accretion depends linearly on the radiative efficiency ϵ_r. The assumption that $\epsilon_r \sim 0.1$ is only justified for nonrotating Schwarzchild black holes [193]. If the black hole is spinning, then its innermost stable circular orbit shrinks [232, 233]. This in turn causes the luminosity and radiative efficiency of the black hole to rise drastically [234], by a factor of up to 4 for a black hole with maximal rotation [21]. Since black holes are spun up by addition of angular

momentum from accretion discs, if growth comes primarily from accretion then the maximum amount of Eddington-limited growth by $z \sim 6$–7 is immensely smaller than calculated in subsection (5.1) above: a factor of $10^{1.9}$ by redshift 7 and $10^{2.1}$ by redshift 6.5, if almost all the growth takes place with the black hole spin $a_* \approx 1$. The only hope to grow $10^9 M_\odot$ black holes in this case would be to assume *much* larger seed black holes or massive amounts of super-Eddington accretion [21]. Of course, if the quasars detected at high redshifts were rapidly spinning, they would have smaller masses than calculated above (80), but only by a linear rather than exponential factor.

If there *is* significant growth from black hole mergers, then significantly lower accretion rates are possible: a simulation by Li *et al.* tracking the growth of the most massive halo in a ~ 3 Gpc^3 volume [21], which treats mergers as effective phases of super-Eddington growth, finds that the overall accretion rate is $\sim 0.1 L_{\rm Edd}$, again assuming $\epsilon_r \sim 0.1$. Significant growth requires efficient merger of black hole binaries, i.e. another source of energy loss besides gravitational wave emission. In a gaseous environment, such as gas-rich early galaxies, strong dynamical friction with the gas can provide the necessary interactions for the binary to decay within ~ 10 Myr [235, 236]. However, if stellar feedback has caused evacuation of the gas surrounding the black holes, merger may take considerably longer [192].

Furthermore, even when the black holes coalesce, the binary can experience strong gravitational recoil in the final stage of merger as a result of a "kick" from anisotropic emission of gravitational waves [237, 238]. If the kick results in a recoil velocity greater than the halo escape velocity, both black holes can be expelled from the halo [239, 240, 241]. Analytic post-Newtonian calculations [242] and fully relativistic numerical simulations [243, 244] of unequal-mass Schwarzchild black holes find a maximum kick velocity of at most $\sim$ 80–300 km $\cdot$ s^{-1}, comfortably below the escape velocities of the merging halos in Li *et al.* [21], 486–1284 km $\cdot$ s^{-1}. However, the situation changes when mergers of spinning black holes are considered: in the extreme case where the merging black holes have maximal, antialigned spins, the kick could be up to 4000 km $\cdot$ s^{-1} [245, 246], although BH spins should become aligned by torques from accreting gas if the binaries are in sufficiently gas-rich environments [247]. Alternatively, high-velocity kicks could be avoided if the mass ratio of the two merging black holes is very small ($M_1/M_2 \ll 10^{-2}$) [248].

5.4 SIDM and Black Holes

Baryons are not the only sort of material that can accrete onto black holes. In [13], Hennawi and Ostriker consider Bondi accretion [171] of self-interacting dark matter onto already-existing central seed black holes. They find that once an optically thick core forms around a central black hole, the speed at which the core can fragment is smaller than the growth rate of the black hole's accretion radius:

$$v_{\rm thermal} \sim \frac{r_c}{r_{\rm thermal}} \sim \frac{v(r_c)}{\tau(r_c)} \ll \frac{dr_c}{dt}. \tag{84}$$

Hence the optically thick region will inescapably be accreted by the black hole, on a timescale independent of the halo scale:

$$t_{\rm Bondi} = \frac{\sigma v_0}{\mu G} = 5.7 \times 10^5 \text{ yr} \left(\frac{\sigma}{1 \text{ cm}^2 \cdot \text{g}^{-1}}\right) \left(\frac{v_0}{100 \text{ km} \cdot \text{s}^{-1}}\right), \tag{85}$$

where μ is defined by equation (53) in subsection 4.1.2 above. Note that, unlike the Salpeter time defined above (82), this is a *total* time, not an e-folding time: SIDM does not generate radiation pressure to resist a black hole's gravitational pull[9]. For galactic-scale values of v_0, the entire optically thick core is accreted in approximately $\sigma/79$ cm$^2 \cdot$ g^{-1} Salpeter times.

Hennawi and Ostriker assumed that the seed black hole had to already be present in the center of the SIDM halo, and considered the conditions for which optically thick regions were formed around the seed black holes which could then be efficiently accreted. However, in the next section we will show that the mechanism of gravothermal collapse in SIDM halos naturally creates both seed black holes and optically thick regions surrounding them.

6 Gravothermal Collapse

We consider the gravothermal collapse of a virialized spherically symmetric ($f(\vec{r},t) = f(r,t)\ \forall f$) dark matter halo, with a mass fraction f of the dark matter interacting via isotropic hard-sphere scattering with a constant cross section per unit mass σ. To avoid excess notation, we will first derive the equations for one species of particle (i.e. $f = 1$) and then generalize

[9] If the χ particles are fermions, it is possible that degeneracy pressure could have the same effect for extremely high densities. An investigation of this possibility is beyond the scope of this paper.

to the fractional case.

6.1 Derivation of the Fundamental Equations

The four fundamental equations governing the behavior of a gravothermal fluid [249] follow from conservation of mass, hydrostatic equilibrium, heat flux, and the laws of thermodynamics. We derive each equation in turn for a general fluid, specializing to SIDM in the next section.

6.1.1 Mass Conservation

For a spherically symmetric mass distribution $M(r)$, conservation of mass states that

$$M(r) = \int_0^r 4\pi r'^2 \rho(r') dr' \ \forall r, \tag{86}$$

where $\rho(r)$ is the density profile. In differential form, this gives the first fundamental equation,

$$\frac{\partial M}{\partial r} = 4\pi r^2 \rho. \tag{87}$$

6.1.2 Hydrostatic Equilibrium

Poisson's equation for the gravitational potential ϕ reads

$$\nabla^2 \phi = 4\pi G \rho; \tag{88}$$

specializing to a spherically symmetric potential and inserting (87) gives

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \phi}{\partial r} \right) = \frac{G \frac{\partial M}{\partial r}}{r^2} \tag{89}$$

which simplifies to

$$\frac{\partial \phi}{\partial r} = \frac{GM}{r^2}. \tag{90}$$

Now apply Newton's Second Law to the case of a non-viscous spherically symmetric fluid

under the influence of a potential ϕ. For any volume V,

$$M\frac{d\vec{v}}{dt} = -\oint_S d^2\vec{S}p(\vec{x}) - M\vec{\nabla}\phi, \tag{91}$$

where S is the surface of V and $p(\vec{x})$ is the pressure exerted on the fluid. But the divergence theorem gives

$$\oint_S d^2\vec{s}\rho = \int_V d^3\vec{x}\vec{\nabla}p \tag{92}$$

so

$$M\frac{d\vec{v}}{dt} = -\int_V dV\vec{\nabla}p - M\vec{\nabla}\phi \rightarrow \int_V \rho dV\frac{d\vec{v}}{dt} = -\int_V dV\left(\vec{\nabla}p - \rho\vec{\nabla}\phi\right) \tag{93}$$

and

$$\rho\frac{d\vec{v}}{dt} = -\vec{\nabla}p - \rho\vec{\nabla}\phi. \tag{94}$$

The total derivative on the left hand side of this equation is a "Lagrangian derivative," measured by an observer travelling along the same path as the fluid. Expanding the velocity change $d\vec{v}$ using the chain rule gives

$$d\vec{v} = \frac{\partial\vec{v}}{\partial t}dt + \sum_{i=1}^{3}\frac{\partial\vec{v}}{\partial x_i}dx_i = \frac{\partial\vec{v}}{\partial t}dt + \left(d\vec{x}\cdot\vec{\nabla}\right)\vec{v} = \frac{\partial\vec{v}}{\partial t}dt + \left(\vec{v}\cdot\vec{\nabla}\right)\vec{v}dt, \tag{95}$$

so

$$\frac{d\vec{v}}{dt} = \frac{\partial\vec{v}}{\partial t} + \left(\vec{v}\cdot\vec{\nabla}\right)\vec{v}. \tag{96}$$

Substitution into (94) gives Euler's equation,

$$\frac{\partial\vec{v}}{\partial t} + \left(\vec{v}\cdot\vec{\nabla}\right)\vec{v} = -\frac{1}{\rho}\vec{\nabla}p - \vec{\nabla}\phi. \tag{97}$$

A fluid in hydrostatic equilibrium has $\vec{v} = 0$, so

$$\vec{\nabla}p = -\rho\vec{\nabla}\phi \rightarrow \frac{\partial p}{\partial r} = -\rho\frac{\partial\phi}{\partial r}. \tag{98}$$

Substituting for ϕ using (90) gives

$$\frac{\partial p}{\partial r} = -\frac{GM\rho}{r^2}. \tag{99}$$

To eliminate p, recall that temperature is defined by

$$\frac{3}{2}k_{\rm B}T = \frac{1}{2}m\left\langle \vec{v}^2 \right\rangle, \tag{100}$$

where m is the mass per particle, and denote the one-dimensional velocity dispersion by ν:

$$\nu^2 \equiv \left\langle v_x^2 \right\rangle = \left\langle v_y^2 \right\rangle = \left\langle v_z^2 \right\rangle = \frac{1}{3}\left\langle \vec{v}^2 \right\rangle = \frac{k_{\rm B}T}{m}. \tag{101}$$

In the case of an ideal gas,

$$p = \frac{\rho}{m}k_{\rm B}T = \rho\nu^2, \tag{102}$$

using the definition of ν^2 above (101). Using this expression for p gives the second fundamental equation,

$$\frac{\partial\left(\rho\nu^2\right)}{\partial r} = -\frac{GM\rho}{r^2}. \tag{103}$$

6.1.3 Heat Flux

Typically the heat flux in a medium is proportional to the temperature gradient: we write

$$q(r) = -\kappa\frac{\partial T}{\partial r}, \tag{104}$$

where q is the radial heat flux density and κ is the thermal conductivity. The luminosity $L(r)$ is the total heat radiated *inward* through a sphere of radius r,

$$L(r) = 4\pi r^2 q(r); \tag{105}$$

eliminating q in favor of L gives the third fundamental equation,

$$\frac{L(r)}{4\pi r^2} = -\kappa\frac{\partial T}{\partial r}. \tag{106}$$

6.1.4 Thermodynamics

The second law of thermodynamics relates heat transfer to changes in entropy,

$$TdS = dQ, \tag{107}$$

or, in terms of specific entropy s,

$$T\left(\frac{\partial s}{\partial t}\right)_M = \frac{1}{\rho V}\left(\frac{dQ}{dt}\right)_M = -\frac{1}{\rho}\vec{\nabla}\cdot\vec{q}, \tag{108}$$

where the negative sign appears because q is the outward heat flux.[10] Specializing to the spherically symmetric case and inserting the luminosity (137) gives

$$T\left(\frac{\partial s}{\partial t}\right)_M = -\frac{1}{4\pi\rho r^2}\frac{\partial L}{\partial r}. \tag{109}$$

On the other hand, the first law of thermodynamics gives

$$dU = TdS - pdV \rightarrow du = Tds + \frac{p}{\rho^2}d\rho, \tag{110}$$

where u and s are the specific energy and entropy, respectively. Solving for ds and using the chain rule to expand du gives

$$ds = \frac{1}{T}\left.\frac{\partial u}{\partial T}\right|_p dT + \left[\frac{1}{T}\left.\frac{\partial u}{\partial p}\right|_T - \frac{p}{T\rho^2}\right]d\rho. \tag{111}$$

In a gas of point particles,

$$u = \frac{3}{2}\frac{k_{\rm B}T}{m}, \tag{112}$$

so

$$ds = \frac{k_{\rm B}}{m}\left[\frac{3}{2}\frac{dT}{T} - \frac{d\rho}{\rho}\right], \tag{113}$$

using $\nu^2 = k_{\rm B}T/m$ (101) and $p = \rho\nu^2$ (102) to simplify the second term. We can now integrate to get

$$s = \frac{k_{\rm B}}{m}\ln\left(\frac{T^{3/2}}{\rho}\right) \tag{114}$$

Inserting (114) into (109) gives

$$\frac{\partial L}{\partial r} = -4\pi\rho r^2\frac{k_{\rm B}T}{m}\left(\frac{\partial}{\partial t}\right)_M \ln\frac{T^{3/2}}{\rho}, \tag{115}$$

[10] We have assumed that there are no mechanisms for internal energy production (such as binary star formation in globular clusters [221, 222], or creation of excited states in the case of inelastic dark matter [165, 166, 167]), so that all heat transfer is due to thermal conduction.

or, again using (101),

$$\frac{\partial L}{\partial r} = -4\pi\rho r^2\nu^2 \left(\frac{\partial}{\partial t}\right)_M \ln\frac{\nu^3}{\rho}, \tag{116}$$

which is the fourth fundamental equation.

6.2 Thermal Conductivity

The four fundamental equations are (87, 103, 106, 116)

$$\frac{\partial M}{\partial r} = 4\pi r^2\rho \tag{117}$$

$$\frac{\partial\left(\rho\nu^2\right)}{\partial r} = -\frac{GM\rho}{r^2} \tag{118}$$

$$\frac{L}{4\pi r^2} = -\kappa\frac{\partial T}{\partial r} \tag{119}$$

$$\frac{\partial L}{\partial r} = -4\pi\rho r^2\nu^2 \left(\frac{\partial}{\partial t}\right)_M \ln\frac{\nu^3}{\rho}. \tag{120}$$

These equations govern the behavior of all monatomic, nonviscous self-gravitating fluids; in order to differentiate SIDM from other gravothermal fluids like globular clusters, we must find an expression for the thermal conductivity κ in terms of σ, our physical parameter.

First we will express the thermal conductivity in terms of quantities from transport theory. Consider the radial heat flux aross a given surface at r. We can write

$$\frac{L(r)}{4\pi r^2} = q(r) = \frac{q_+(r) + q_-(r)}{2}; \tag{121}$$

i.e. the net flow of heat is equal to the sum of the outward-moving and inward-moving heat flux. But the heat flux is just the particle flux multiplied by the thermal energy per particle:

$$q(r) = \frac{1}{2}\left[-\left(\frac{n\lambda}{\tau}\right)\frac{3}{2}k_{\mathrm{B}}T_+ + \left(\frac{n\lambda}{\tau}\right)\frac{3}{2}k_{\mathrm{B}}T_-\right], \tag{122}$$

with λ the mean free path, τ the collision time, and where T_+ is the temperature just above the surface and T_- is the temperature just below it, and we have assumed a monatomic gas

with equipartition of energy between the particles. To lowest order,

$$T_{\pm} = T(r) \pm b\lambda \frac{\partial T}{\partial r}, \tag{123}$$

where b is an effective impact parameter, which can be calculated perturbatively using Chapman-Enskog theory[11] [250, 251], $b = 25\sqrt{\pi}/32 \approx 1.385$, so

$$\frac{L}{4\pi r^2} = -\frac{3}{2} b\rho \frac{\lambda^2}{\tau} \frac{\partial \nu^2}{\partial r}, \tag{124}$$

where we have used the relation between T and ν^2 (101) to simplify the expression. This is a general result for monatomic fluids.

To determine the values of λ and τ for SIDM, we will first consider two limiting cases and then interpolate between them to find a general result. Behavior of an SIDM fluid depends critically on the relative magnitudes of the mean free path λ and the typical radial distance, which is the Jeans length or gravitational scale height H:

$$H = \frac{1}{k_{\rm J}},\ k_{\rm J}^2 \nu^2 = 4\pi G\rho \rightarrow H = \sqrt{\frac{\nu^2}{4\pi G\rho}}. \tag{125}$$

In the short mean free path limit, $\lambda \ll H$, the usual relation $\tau = \lambda/\nu$ applies, and

$$\frac{L}{4\pi r^2} = -\frac{3}{2} b\rho\nu\lambda \frac{\partial \nu^2}{\partial r}. \tag{126}$$

In the long mean free path limit, $\lambda \gg H$, however, we must take $\lambda \rightarrow H$ and $\tau \rightarrow t_r$, the relaxation time (we will justify using t_r rather than the dynamic timescale $t_d \equiv H/\nu$ in footnote 12 below). Then

$$\frac{L}{4\pi r^2} = -\frac{3}{2} b\rho \frac{H^2}{t_r} \frac{\partial \nu^2}{\partial r}. \tag{127}$$

We can combine the two expressions together in reciprocal to form a general expression

[11] Balberg, Shapiro, and Inagaki [23] erroneously quote $b = 25\pi/\left(32\sqrt{6}\right) \approx 1.002$. The correct result is given in Koda and Shapiro [24].

with the correct limiting behavior:

$$\frac{L}{4\pi r^2} = -\frac{3}{2} b\rho\nu \left[\frac{1}{\lambda} + \frac{b}{C}\frac{\nu t_r}{H^2}\right]^{-1} \frac{\partial \nu^2}{\partial r}, \tag{128}$$

where C is a constant which sets the scale on which the two mechanisms are equally effective. C cannot be determined analytically, but comparison to the results of N-body SIDM simulations [24] gives $C \approx 290/385 \approx 0.753$. Finally, the mean free path is given by $\lambda = 1/(\rho\sigma)$, while, since we have assumed close-range interactions, the relaxation time is equal to the mean time between collisions[12]:

$$t_r = \frac{1}{a\rho\nu\sigma}, \tag{129}$$

where a is a constant of order unity, $a = \sqrt{16/\pi} \approx 2.257$ for hard-sphere interactions [252]. Substituting for λ and $t_{\rm r}$ yields the final form of the equation for $f = 1$:

$$\frac{L}{4\pi r^2} = -\frac{3}{2} ab\nu\sigma \left[a\sigma^2 + \frac{b}{C}\frac{4\pi G}{\rho\nu^2}\right]^{-1} \frac{\partial \nu^2}{\partial r}. \tag{130}$$

6.3 Fractional SIDM

When $f = 1$, the fundamental equations (87, 103, 130, 116) are a set of four partial differential equations with four dependent variables $\{M,\ \rho,\ \nu,\ L\}$ and two independent variables $\{r,\ t\}$. In the fractional case, however, there are two species of dark matter, interacting and non-interacting. Because hydrostatic equillibrium is satisfied for each species of particle independently[13], the two species have independent profiles, subject to the constraint

$$\frac{\partial M}{\partial r} = 4\pi r^2 \left(\rho^{int} + \rho^{ni}\right), \tag{131}$$

[12] We see that $\lambda \gg H$ implies $t_r \gg t_d$, so our use of t_r in the long mean free path limit is justified. This implies that the timescale to return to hydrostatic equillibrium is much faster than the conduction timescale, which we will use in subsection 6.4.2 below.

[13] This follows from Dalton's law: since the two species do not interact at short ranges with each other, the total pressure must be equal to the sum of the partial pressures. Hence equillibrium is only possible for the fluid as a while when both species are individually in equillibrium.

where ρ^{int} and ρ^{ni} represent the interacting and non-interacting density profiles, respectively, and conservation of mass requires that

$$\frac{\int_0^\infty 4\pi r'^2 \rho^{int}(r')dr'}{\int_0^\infty 4\pi r'^2 \rho^{ni}(r')dr'} = \frac{f}{1-f}. \tag{132}$$

Since $\sigma = 0$ for non-interacting particles, 130 is trivially satisfied by

$$L^{ni}(r) = 0 \ \forall r. \tag{133}$$

Hence the total system consists of six partial differential equations with six dependent variables $\{M,\ \rho^{int},\ \rho^{ni},\ \nu^{int},\ \nu^{ni},\ L^{int}\}$ and two independent variables $\{r,\ t\}$:

$$\frac{\partial M}{\partial r} = 4\pi r^2 \left(\rho^{int} + \rho^{ni}\right) \tag{134}$$

$$\frac{\partial \left(\rho^{int}\left(\nu^{int}\right)^2\right)}{\partial r} = -\frac{GM\rho^{int}}{r^2} \tag{135}$$

$$\frac{\partial \left(\rho^{ni}\left(\nu^{ni}\right)^2\right)}{\partial r} = -\frac{GM\rho^{ni}}{r^2} \tag{136}$$

$$\frac{L^{int}}{4\pi r^2} = -\frac{3}{2}ab\nu^{int}\sigma\left[a\sigma^2 + \frac{b}{C}\frac{4\pi G}{\rho^{int}\left(\nu^{int}\right)^2}\right]^{-1}\frac{\partial\left(\nu^{int}\right)^2}{\partial r} \tag{137}$$

$$\frac{\partial L^{int}}{\partial r} = -4\pi\rho^{int}r^2\left(\nu^{int}\right)^2\left(\frac{\partial}{\partial t}\right)_M \ln\frac{\left(\nu^{int}\right)^3}{\rho^{int}} \tag{138}$$

$$0 = \left(\frac{\partial}{\partial t}\right)_M \ln\frac{\left(\nu^{ni}\right)^3}{\rho^{ni}}. \tag{139}$$

The first equation allows the total mass distribution to be calculated from the individual density distributions. The second and third equations enforce hydrostatic equilibrium for each species, and, through the factor of M, represent the only coupling between the two species. The fourth and fifth equations govern SIDM conduction, while the last tells us that the entropy of the collisionless CDM is conserved,

$$3\frac{\dot{\nu}}{\nu} = \frac{\dot{\rho}}{\rho}. \tag{140}$$

In principle, the six master equations (134 –139) can be solved exactly given appropriate boundary conditions at $r = 0$ and $r = \infty$ and a set of initial radial profiles which obey the equations. However, in practice this is computationally impossible. Instead, the strategy followed by [23, 24], which we will adopt, is to start from a previously motivated set of radial profiles, e.g. an NFW density profile, and numerically integrate forward in time.

6.4 Numerical Integration

6.4.1 Dimensionless Form of the Equations

In order to do numerical work, it is convenient to switch to a dimensionless form of the equations. Choose fiducial mass and length scales, M_0 and R_0. Then there are natural scales for the remaining dependent variables [17]:

$$\rho_0 = \frac{M_0}{4\pi R_0^3},\ \nu_0 = \sqrt{\frac{GM_0}{R_0}},\ L_0 = \frac{GM_0}{R_0 t_0}, \tag{141}$$

Given this choice of scales, the interaction cross section can be written

$$\sigma = \hat{\sigma}\sigma_0,\ \sigma_0 = \frac{4\pi R_0^2}{M_0}, \tag{142}$$

and the characteristic time scale is given by

$$t_0 = \frac{R_0}{ab\nu_0\hat{\sigma}} = \frac{t_{r,0}}{b\hat{\sigma}}, \tag{143}$$

where $t_{r,0}$ denotes the relaxation time calculated from the fiducial quantities. Then define dimensionless variables by $\tilde{x} \equiv x/x_0$, i.e. $\tilde{r} = r/r_0$, $\tilde{\rho^{int}} = \rho^{int}/\rho_0$. The dimensionful quantities can be restored by specifying the physical values of σ and $t_{r,c}(0)$, the initial relaxation time at the center of the profile. This choice of variables simplifes the structure of the equations by removing the factors of 4π and G everywhere they appear, and replacing all variables by their dimensionless equivalents (and σ by $\hat{\sigma}$).

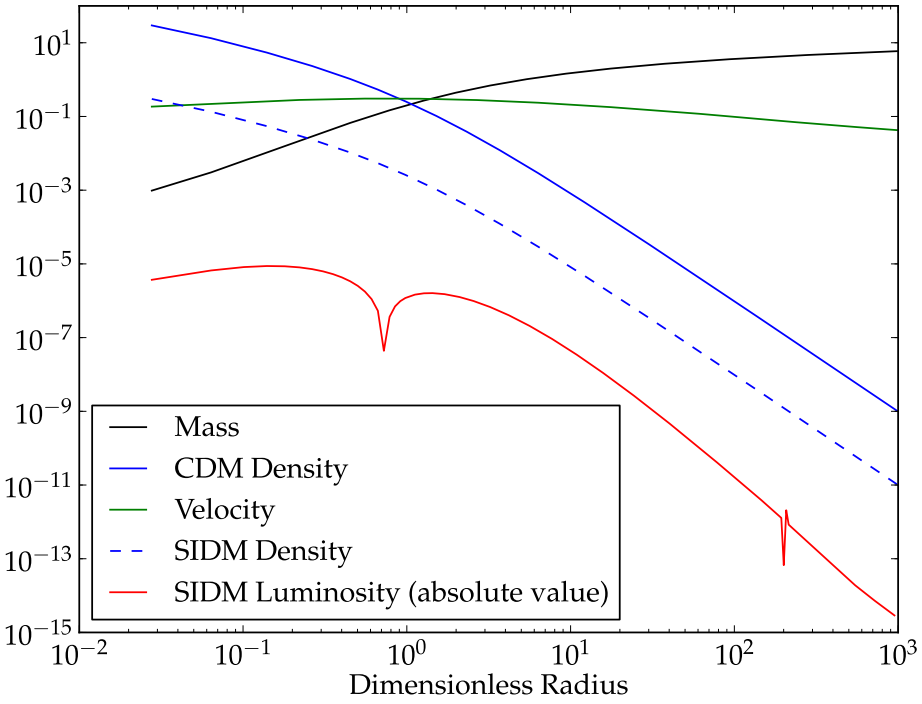

Figure 6: The dimensionless initial profiles for an NFW halo, with $f = 0.01$. The CDM and SIDM have the same velocity profile, and their density profiles have the same shape but a normalization differing by $f/(1+f)$. As expected for SIDM, the initial luminosity at small radii is negative, indicating that the cusp is being forced outward as a core begins to form. The lurch in the luminosity at $\tilde{r} = 200$ is a numerical artifact.

6.4.2 Integration Strategy

Following [23, 24], we spatially discretize the problem, starting with an array of N spherical shells which are initially evenly logarithmically spaced in radius. We will use $\tilde{r}_i$ to denote the outer radius of the ith shell at a given timestep, and abbreviate $\tilde{M}(\tilde{r}_i) = \tilde{M}_i$, etc. Extensive quantities (e.g. $\tilde{M}_i$, $\tilde{V}_i$, $\tilde{L}_i$) denote the value at/enclosed within $\tilde{r}_i$, while intensive quantities (e.g. $\tilde{\rho}_i$, $\tilde{\nu}_i$, $\tilde{u}_i$) denote the average value in the ith shell, i.e. the value at $(\tilde{r}_{i-1} + \tilde{r}_i)/2$, since we are assuming the spacing between shells is very small. Expressions with no superscripts should be understood as applying to both the interacting and non-interacting quantities.

As in section 4 above, we start with both the interacting and non-interacting components

in an NFW profile (27), and choose the fiducial radius $R_0 = r_s$, so in dimensionless units (where the usual factors of 4π that appear in spherical integration are dropped)

$$\tilde{M}_i = \int_0^{\tilde{r}} \tilde{\rho}\tilde{r}^2 d\tilde{r} = \int_0^{\tilde{r}_i} \tilde{r}^{-1}\left(1+\tilde{r}\right)^{-2}\tilde{r}^2 d\tilde{r} = -\frac{\tilde{r}}{1+\tilde{r}} + \ln(1+\tilde{r}). \tag{144}$$

The dimensionless density is then given by[14]

$$\tilde{\rho}_i = 3\frac{\tilde{M}_i - \tilde{M}_{i-1}}{\tilde{V}_i - \tilde{V}_{i-1}} = 3\frac{\tilde{M}_i - \tilde{M}_{i-1}}{\tilde{r}_i^3 - \tilde{r}_{i-1}}, \tag{145}$$

and the densities of the SIDM and CDM are a factor f and $1-f$, respectively, of this total density. Instead of directly manipulating the velocity and luminosity, it is convenient to track the dimensionless pressure (102) $\tilde{p} = \tilde{\rho}\tilde{\nu}^2$ and specific energy (101,112) $\tilde{u} = 3\tilde{\nu}^2/2$. The initial values $\tilde{p}_i$ are given by integrating the dimensionless force per volume:

$$\tilde{p}_i = \int_{(\tilde{r}_i+\tilde{r}_{i+1})/2}^{\infty} \tilde{\rho}\tilde{M} d\tilde{r}, \tag{146}$$

and $\tilde{u}_i = 3\tilde{p}_i/(2\tilde{\rho}_i)$. Figure 6 shows the initial mass, density, and velocity profiles for the CDM and SIDM, as well as the SIDM luminosity profile.

Given the initial conditions, we can begin to numerically integrate the profiles forward in time. Since the relaxation time is much longer than the dynamical time, as discussed in footnote 12, we are justified in taking a small conduction/diffusion timestep and then instantaneously adjusting the profiles to maintain hydrostatic equilibrium while keeping entropy constant. To simulate the effects of conduction, we first calculate the SIDM luminosity profile using the dimensionless, discretized form of (137):

$$\tilde{L}_i^{int} = -\tilde{r}_i^2 \frac{\sqrt{2\tilde{u}_i^{int}/3} + \sqrt{2\tilde{u}_{i+1}^{int}/3}}{2}\left[\frac{1}{Ca}\left(\frac{\tilde{p}_i^{int}+\tilde{p}_{i+1}^{int}}{2}\right)^{-1} + \frac{\hat{\sigma}^2}{b}\right]^{-1}\frac{\tilde{u}_{i+1}^{int}-\tilde{u}_i^{int}}{\tilde{r}_{i+1}-\tilde{r_{i-1}}}, \tag{147}$$

where the last term is a discrete derivative, the factor of 3/2 in (137) was absorbed because we are using $\tilde{u}$ instead of $\tilde{\nu}^2$, and the averaging of $\tilde{v}$ and $\tilde{p}$ is because of the conventions for

[14] Clearly this is not the value of $\tilde{\rho}_0$—we need to make the implicit assumption that $\tilde{M}_{-1} = \tilde{r}_{-1} = 0$. In general, the discretized expressions $\tilde{x}_i$ in this section will need to be modified in a similar fashion at the endpoints, i.e. for $\tilde{x}_0$ and/or $\tilde{x}_N$.

intensive vs. extensive quantities defined in the first paragraph of this subsection.

Once the luminosity profile has been calculated, the energy flowing into a shell during a given timestep $\Delta\tilde{t}$ is distributed across the SIDM in that shell:

$$\Delta\tilde{u}_i = -\frac{\tilde{L}_i^{int} - \tilde{L}_{i-1}^{int}}{\tilde{M}_i^{int} - \tilde{M}_{i-1}^{int}}\Delta\tilde{t}, \tag{148}$$

where $\tilde{M}_i^{int}$ denotes the SIDM mass in the ith shell,

$$\tilde{M}_i^{int} = M_{int}\frac{\tilde{\rho}_i^{int}}{\tilde{\rho}_i^{int} + \tilde{\rho}_i^{ni}}. \tag{149}$$

The timestep is chosen to be a small portion of the fastest relaxation time in the profile, so that $\Delta\tilde{u}_i/\tilde{u}_i \ll 1\ \forall i$.

Following the conduction/diffusion timestep, the SIDM and CDM profiles are adjusted to maintain hydrostatic equilibrium. Since this process is entropy-preserving, there is an adiabatic invariant A_i for each shell,

$$A_i \sim \tilde{p}_i\tilde{V}_i^{\,5/3} \sim \frac{2}{3}\tilde{\rho}_i\tilde{e}_i\left(\frac{\tilde{M}_i}{\tilde{\rho}_i}\right)^{5/3} \sim \frac{2}{3}\tilde{\rho}_i^{-2/3}\tilde{e}_i, \tag{150}$$

since we are using the method of Lagrangian zones [23] and adjusting the radius but keeping the mass fixed to maintain hydrostatic equilibrium. The conduction/diffusion timestep will have moved each shell out of hydrostatic equilibrium (135),

$$\Delta_i = \frac{\tilde{p}_{i+1} - \tilde{p}_i}{\left(\tilde{r}_{i+1} - \tilde{r}_{i-1}\right)/2} + \frac{\tilde{\rho}_i + \tilde{\rho}_{i+1}}{2}\frac{\tilde{M}_i}{\tilde{r}_i^2} \neq 0, \tag{151}$$

where in practice we use $A_i\tilde{\rho}_i^{5/3} = \tilde{\rho}_i\tilde{e}_i/1.5 = \tilde{p}_i$ to find the new value of $\tilde{p}_i$ after the conduction/diffusion timestep. The problem is to find a set of $\Delta\tilde{r}_i$ which alter the pressure, radius, and density of each shell such that hydrostatic equilibrium is again satisfied:

$$\frac{\tilde{p}_{i+1} + \Delta\tilde{p}_{i+1} - \tilde{p}_i - \Delta\tilde{p}_i}{\left(\tilde{r}_{i+1} + \Delta\tilde{r}_{i+1} - \tilde{r}_{i-1} - \Delta\tilde{r}_{i-1}\right)/2} + \frac{\tilde{\rho}_i + \Delta\tilde{\rho}_i + \tilde{\rho}_{i+1} + \Delta\tilde{\rho}_{i+1}}{2}\frac{\tilde{M}_i}{\tilde{r}_i^2 + 2\Delta\tilde{r}_i} = 0. \tag{152}$$

First compute the change in dimensionless volume between adjacent shells:

$$\Delta\tilde{V}_i = \tilde{r}_i^2 \Delta\tilde{r}_i - \tilde{r}_{i-1}^2 \Delta\tilde{r}_{i-1}. \tag{153}$$

Since we have assumed that the mass of each shell is unchanged, we must have

$$\Delta\tilde{\rho}_i = -\tilde{\rho}_i \frac{\Delta\tilde{V}_i}{\tilde{V}_i} = -3\tilde{\rho}_i \frac{\tilde{r}_i^2 \Delta\tilde{r}_i - \tilde{r}_{i-1}^2 \Delta\tilde{r}_{i-1}}{\tilde{r}_i^3 - \tilde{r}_{i-1}^3}. \tag{154}$$

Similarly,

$$\Delta\tilde{p}_i = -\frac{5}{3}\tilde{p}_i \frac{\Delta\tilde{V}_i}{\tilde{V}_i} = -5\tilde{p}_i \frac{\tilde{r}_i^2 \Delta\tilde{r}_i - \tilde{r}_{i-1}^2 \Delta\tilde{r}_{i-1}}{\tilde{r}_i^3 - \tilde{r}_{i-1}^3}. \tag{155}$$

Inserting these changes into the hydrostatic equilibrium equation (152), multiplying the expression by both denominators, and linearizing in the $\Delta\tilde{r}_i$s results in a tridiagonal system of equations, which can be solved using a standard linear algebra library [253]. Once hydrostatic equilibrium is reached, the specific energies $\tilde{u}_i$ are also adjusted so that A_i is constant (150). Another conduction/diffusion timestep can then be taken, and in principle the forward integration can be continued indefinitely. In practice, when core collapse occurs the densities rise sufficiently quickly that the integration timestep becomes extremely small.

6.4.3 A Problem

Although the procedure just described seems like it should work just as well for multiple species as it does for one, we have glossed over a difficulty in the previous subsection. In order to compute the deviation from hydrostatic equilibrium (151), we needed to evaluate $\tilde{M}_i$—i.e. $\tilde{M}(\tilde{r}_i)$, the total mass enclosed within a sphere with radius $\tilde{r}_i$. We managed to calculate $\tilde{\rho}_i$ from the initial mass distribution (145), but we cannot invert this equation and build up $\tilde{M}_i$ iteratively,

$$\tilde{M}_i = \frac{1}{3}\left(\tilde{\rho}_i^{int} + \tilde{\rho}_i^{ni}\right)\left(\tilde{r}_i^3 - \tilde{r}_{i-1}^3\right) + \tilde{M}_{i-1}, \tag{156}$$

because *the sets of radii $\tilde{r}_i$ are different for the SIDM and CDM.* Because the SIDM gains energy from conduction while the CDM does not, the two species will relax into hydrostatic equilibrium differently. In principle, this problem can be solved by interpolating the density

distribution to approximate the value of $\tilde{\rho}$ at any $\tilde{r}_i$. However, for computationally feasible numbers of shells, $N \sim 200$, the error introduced by e.g. cubic splines rapidly becomes large enough to overwhelm the true results.

More work clearly remains to be done on ensuring numerical convergence over many relaxation times. However, the existing level of stability, which can be maintained over $\mathcal{O}(10)$ SIDM initial relaxation times, is enough to suggest an interesting result. Figure 7 plots the evolution of the SIDM density profile for an $f = 0.01$ halo starting from an initial NFW profile. As expected, the profile begins to flatten and develop a core. Stability issues and computational resources prevent integrating this fractional halo until collapse. However, it seems that the profile is flattening approximately as occurs for an $f = 1$ NFW halo (see the next subsection), although on longer timescales because the relaxation time goes as f^{-1}. This is not unexpected: the conduction/diffusion timestep takes the system out of equillibrium by increasing the energy and therefore the pressure in a shell, i.e. by increasing the *first* term in (151). Hence we should expect this term to be much larger in magnitude than the second, mass term, implying that adjustments from interactions with the non-interacting component are unimportant compared to the conduction itself.

6.4.4 Non-Fractional NFW Profile Evolution

Figure 8 shows the density evolution of an $f = 1$ halo starting from NFW initial conditions. The first ~ 25 relaxation times are shown: beyond this point, the density continues to shrink but at an increasingly slow rate. As in the $f = 0.01$ case, a core develops as heat conduction dissolves the initial cusp. Careful comparison of Figure 8 and the $f = 0.01$ case in Figure 7 above reveals that the presence of additional non-interacting dark matter makes a quantitative difference, though not a qualitative one, to the SIDM's evolution. Schematically, the relative initial mass loss is greater in the $f = 0.01$ case, but subsequent evolution proceeds more slowly (in terms of relaxation times, which of course are significantly longer in the $f = 0.01$ case). This is heuristically what we would expect: initial mass loss is easier because there is less total SIDM mass in the central regions, but evolution is slowed because of interaction with the still-cuspy CDM halo.

Figure 9 plots the mass, density, velocity, and luminosity profiles of the $f = 1$ halo after 23.39 relaxation times, when the density at $\tilde{r} = 10^{-2}$ has shrunk by over an order of

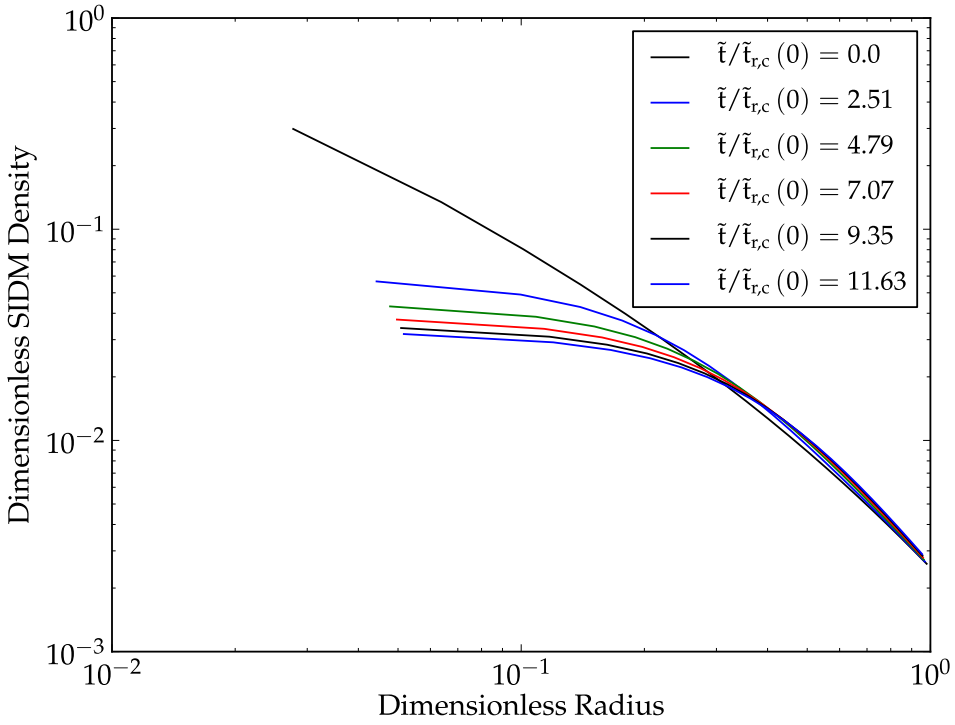

Figure 7: Evolution of SIDM density profiles, starting with an $f = 0.01$ NFW halo. Only the inner part of the halo is shown; the outer part still asymptotes to r^{-3} as in Figure 6 above for all halos. From top to bottom, profiles are at 0.0, 2.51, 4.79, 7.07, 9.35, and 11.63 central relaxation times. Because $t_r \propto \rho^{-1}$, this corresponds to integrating for ~ 1000 relaxation times in an $f = 1$ halo. However, comparison to the $f = 1$ results in Figure 8 below suggests that the density profile flattens in the same manner, just f^{-1} times slower: evidently the non-interacting dark matter has little influence on the central SIDM evolution.

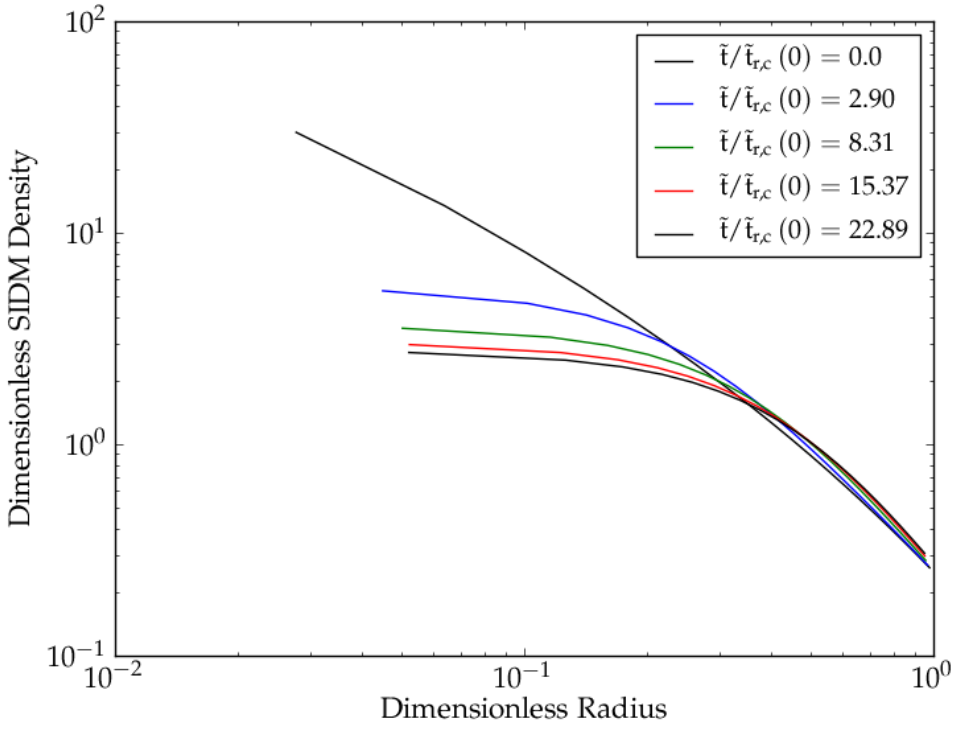

Figure 8: Evolution of an $f = 1$ halo starting from NFW initial conditions. For clarity, only the inner portion of the density profile is shown: the outer profile has not yet changed significantly at this stage. From top to bottom, profiles are at 0.0, 2.90, 8.31, 15.37, and 22.89 central relaxation times. As in the $f = 0.01$ case, the density profile is flattening as a core develops.

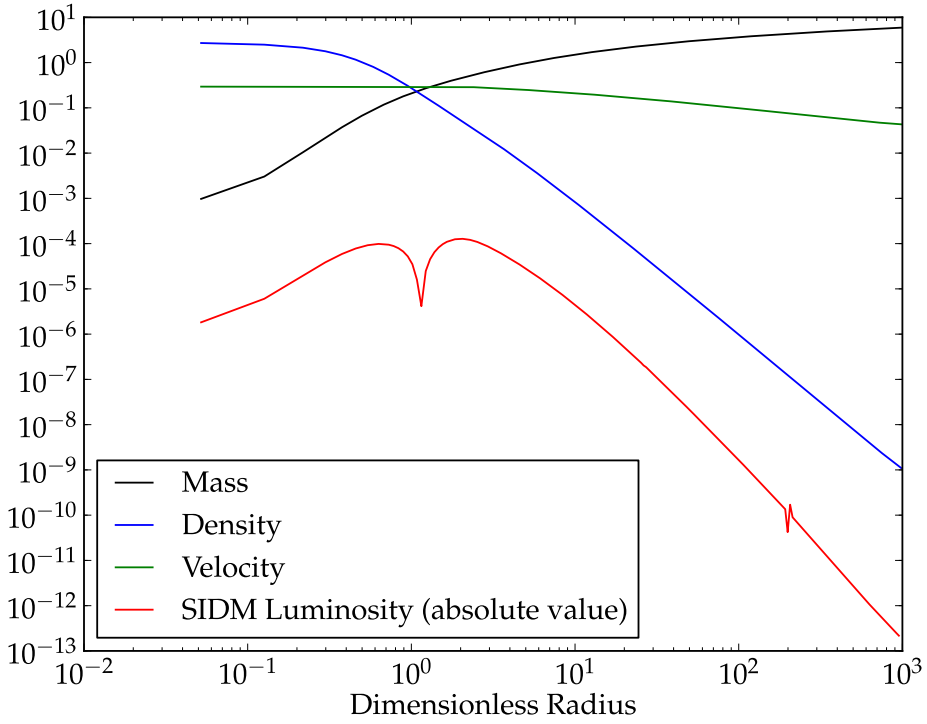

Figure 9: Dimensionless profiles for an $f = 1$ SIDM halo, 23.39 central relaxation times after starting with NFW initial conditions. We see that the density profile is leveling out, the velocity profile is becoming flat in the emerging core, and the amplitude of the luminosity is rising at large radii as more mass is transferred there. Evidently the profiles are gradually evolving towards the self-similar profile in Figure 10 below. As in Figure 6 above, the glitch in the luminosity near $\tilde{r} = 200$ is a numerical artifact.

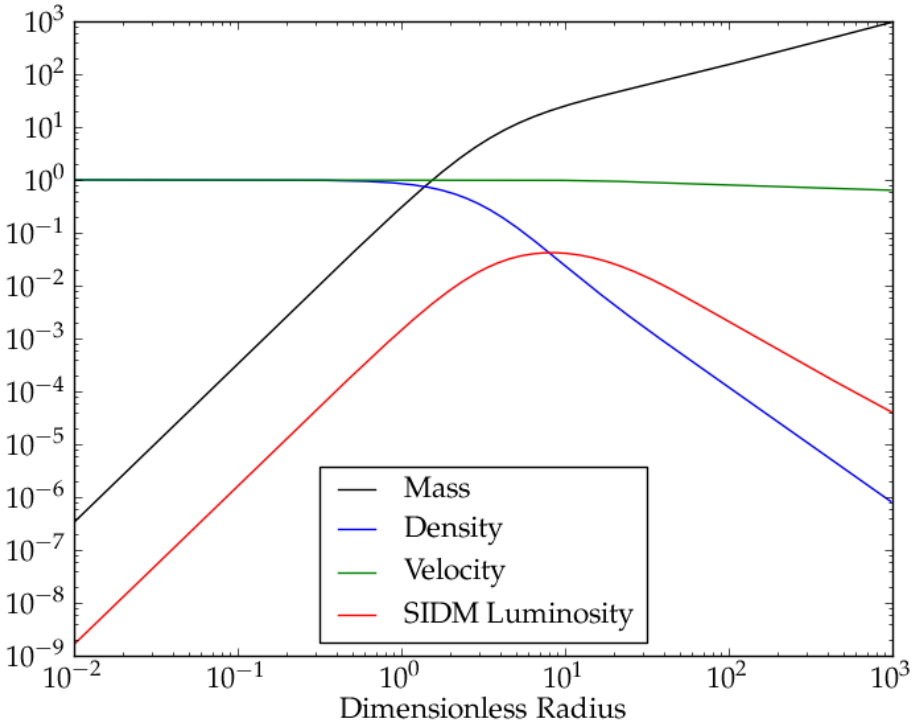

Figure 10: Initial conditions for the self-similar SIDM halo. The mass and velocity profiles have become completely flat below $\tilde{r} = 1$, while the luminosity has increased by another two orders of magnitude, peaked around $\tilde{r} = 10$, and become positive everywhere, indicating that the core density will not shrink further.

magnitude. We can clearly see that loss of mass from the central reason is causing a core to develop, leveling the velocity profile and dumping mass onto the r^{-3} tail, increasing the luminosity in the outer regions.

6.4.5 The Self-Similar Profile

According to the simulations of Koda and Shapiro [24], after $\mathcal{O}(100)$ relaxation times the $f = 1$ halo will have completed development of a flat core with $\tilde{\rho} = 1$, extending out to $\tilde{r}_c = 1$. At this point, at least in the limit of small $\hat{\sigma}$, the SIDM halo can now be described by a *self-similar* solution, where further time evolution does not change the shape of the density profile, only its amplitude. Figure 10 shows the mass, density, velocity, and

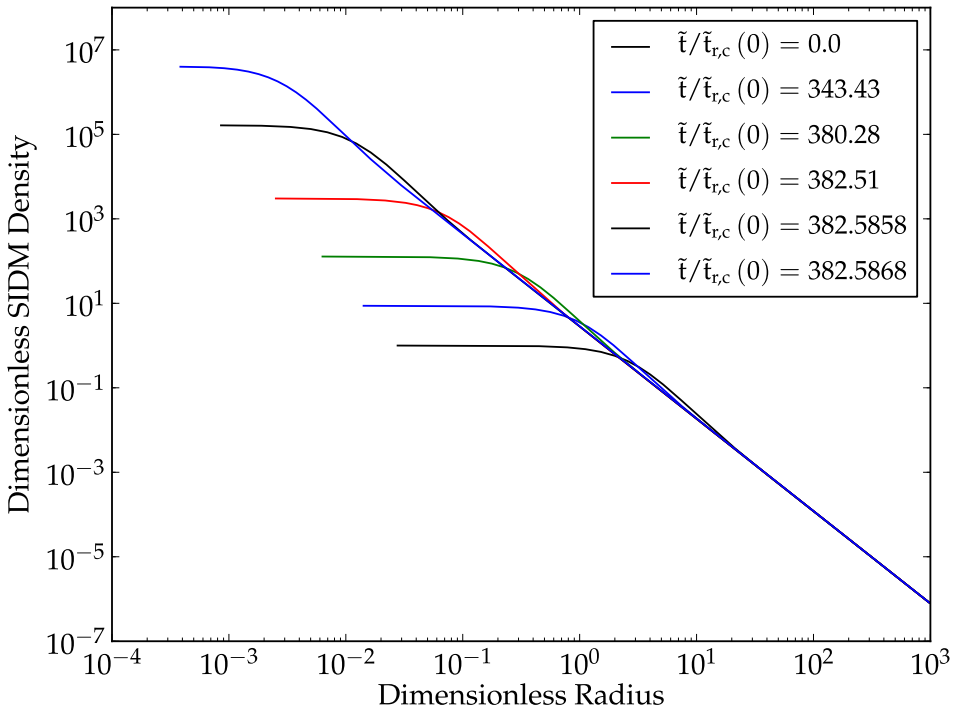

Figure 11: Runaway collapse of a SIDM halo in the $\hat{\sigma} \ll 1$ limit, starting from a self-similar profile. From bottom to top, profiles are at 0.0, 343.43, 380.28, 382.51, 382.5858, and 382.5868 central relaxation times after the self-similar profile initially forms. The density profile remains self-similar and increases in density catastrophically as $\tilde{t}/\tilde{t}_{r,c}(0)$ approaches ~ 382.59.

luminosity profiles of a self-similar halo with $\tilde{r}_c = 1$. The trends evident in Figure 9 above, at an intermediate stage in the evolution from NFW initial conditions, have continued: the density and velocity profiles are flat at large distances, while the luminosity profile has become everywhere positive, preventing the core density from shrinking further.

Once the core is stable, the everywhere-positive luminosity will slowly cause the core to become smaller and denser. Eventually this will lead to runaway core collapse: unlike the case of globular clusters, (elastic) SIDM halos have no equivalent of binary star formation to halt gravothermal collapse. When velocities become relativistic, the core will undergo catastrophic collapse into a black hole.

In the limit of small $\hat{\sigma}$, the evolution will be self-similar, and the density profile will look like the initial profile in Figure 10 throughout. Figure 11 confirms this self-similar evolution in the $\hat{\sigma} \ll 1$ limit. The core shrinks and rises in density catastrophically; approximately 383 relaxation times after the initial self-similar profile forms, it will inevitably undergo collapse into black hole. Even if only a portion of the inner core initially collapses into a black hole, the remainder will remain in the short mean free path limit, so, as discussed in subsection 5.4 above, Bondi accretion will rapidly grow the black hole to the point where it encompasses the entire inner core.

However, if $\hat{\sigma}$ is non-negligible, as shown in Figure 12 for the case of $\hat{\sigma} = 0.088$, the core eventually becomes so dense that it enters the short mean free path regime and becomes optically thick. When this happens, the $a\hat{\sigma}^2$ term in the luminosity equation (137) becomes non-negligible, introducing a scale into the problem. Accordingly, the profile is no longer self-similar: while a low-mass inner core continues the runaway collapse process, an outer core also develops to transition between the inner core and the outer regions. In addition, a larger $\hat{\sigma}$ slows the core collapse process slightly, since the short mean free path conductivity (126) is proportional to $\lambda \propto \sigma^{-1}$.

Figure 13 shows the evolution of the mass profile for the same $\hat{\sigma} = 0.088$ halo. We see that once an inner core develops in the short mean free path regime, it contains a nearly constant mass, $M_c \approx 2.5 \times 10^{-3} M_{\rm tot}$. In the case of a fractional SIDM halo, we expect that the inner core will contain the same fraction of the total SIDM in the halo,

$$M_c \approx 2.5 \times 10^{-3} f M_{tot}. \tag{157}$$

As discussed above, this will also be the mass of the seed black hole after catastrophic collapse and Bondi accretion of the optically thick areas.

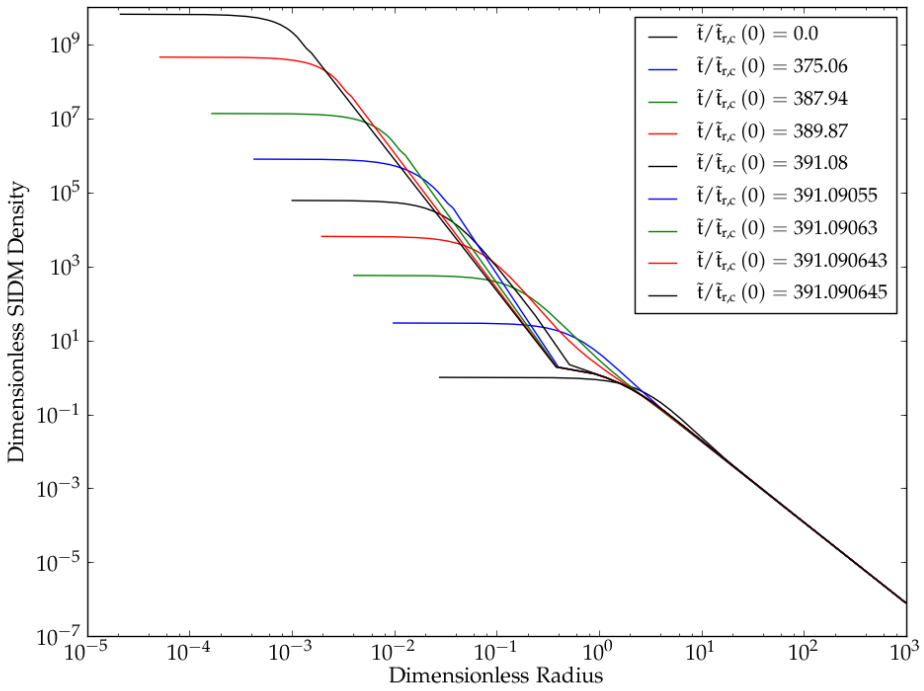

Figure 12: Runaway collapse of a cored SIDM halo with $\hat{\sigma} = 0.088$, starting from a self-similar profile. From bottom to top, profiles are at 0.0, 375.06, 387.94, 389.87, 391.08, 391.09055, 391.09063, 391.090643, and 391.090645 central relaxation times after the self-similar profile initially forms. After ~ 390 relaxation times, the core of the halo becomes optically thick, and self-similarity is broken: the core splits into a very dense inner core and an outer core which transitions between the two regions. Nevertheless, catastrophic collapse occurs as $\tilde{t}/\tilde{t}_{r,c}(0)$ approaches ~ 391.09.

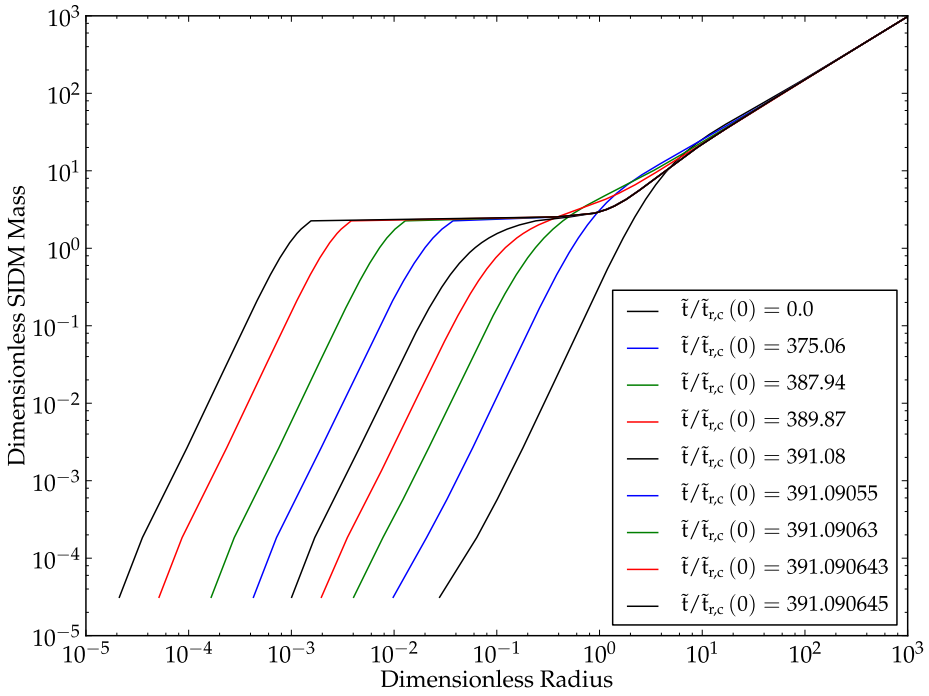

Figure 13: Mass profile history of a cored SIDM halo with $\hat{\sigma} = 0.088$, starting from a self-similar profile. From right to left, profiles are at 0.0, 375.06, 387.94, 389.87, 391.08, 391.09055, 391.09063, 391.090643, and 391.090645 central relaxation times after the self-similar profile initially forms. Once the core enters the optically thick regime, around $\tilde{t}/\tilde{t}_{r,c}(0) = 390$, the inner core contains a constant total mass, around 2.5×10^{-3} of the total SIDM mass.

7 SIDM and SMBHs

In Section 5, we saw that Bondi accretion of self-interacting dark matter onto seed black holes allows highly super-Eddington growth, which could alleviate the tensions in the standard explanation of $10^9 M_\odot$ black holes at high redshifts. Subsequently, in Section 6 we saw that gravothermal collapse of fractional SIDM halos not only creates the optically thick conditions required for Bondi accretion, but provides the central seed black hole in addition. In this section, we combine all of the pieces to present a SIDM-driven scenario which can account for the observed spatial density of high-z SMBHs.

7.1 Cross Section

In subsection 6.4.5 above, we saw that $f = 1$ self-similar SIDM halos collapse in ~ 390 central relaxation times, depending on the precise value of $\hat{\sigma}$, and form a central black hole with mass $\sim 2.5 \times 10^{-3} M_{\rm tot}$. In the previous subsection, we hypothesized that initial NFW halos first lose central mass, form a core, and develop into the self-similar profile before undergoing core collapse. Following [24], we assume that the time to develop into a self-similar halo is ~ 100 relaxation times. Finally, we claimed that fractional SIDM halos should also develop self-similar profiles, although perhaps on a slightly longer timescale. For the remainder of this section, we will therefore assume that[15]

$$t_{\rm collapse} = 490 t_r. \tag{158}$$

Hence in order for a halo formed at redshift z_0 to have undergone core collapse before a given redshift z_1, we need

$$t(z_1) - t(z_0) \geq 490 t_r, \tag{159}$$

where $t(z)$ is the time after the Big Bang corresponding to redshift z,

$$t(z) = t_0 \int_0^{1/(1+z)} \frac{da}{\dot{a}} = t_0 \int_0^{1/(1+z)} \frac{da}{aH} = t_0 \int_0^{1/(1+z)} \frac{da}{a\sqrt{\Omega_\gamma a^{-2} + \Omega_m a^{-3} + \Omega_\Lambda}}, \tag{160}$$

[15] Because the slope of the halo mass function is very steep at high redshifts, altering the coefficient by a small amount will significantly change the (σ, $M_{\rm tot}$, f) values derived below. However, we emphasize that there is nothing special about the value chosen: much larger coefficients can still be accomodated by altering the assumed value of f or the amount of growth that comes from Eddington-limited gas accretion.

where t_0 is the current age of the universe. Neglecting the energy in radiation and setting [1] $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, $t_0 = 13.75$ Gyr, gives the following numerical approximation, accurate when the energy density in radiation Ω_γ is negligible:

$$t(z) = 0.78 t_0 \sinh^{-1}\left(1.64\left(\frac{1}{1+z}\right)^{3/2}\right). \tag{161}$$

Our goal in the remainder of this subsection, following [17], is to exploit the relation (159) to derive a lower bound on the cross section per unit mass σ necessary for a halo of mass $M_{\rm tot}$ with SIDM fraction f formed at redshift z_0 to undergo gravothermal collapse by redshift z_1.

First, recall the expression for relaxation time (129),

$$t_r(0) = \frac{1}{af\rho_c(0)\nu_c(0)\sigma}. \tag{162}$$

In order to express ρ_c and ν_c in terms of the halo parameters, define

$$x \equiv \frac{\rho_c(0)}{\rho_{\rm crit}(z_0)},\ g(x) \equiv \frac{M_{\rm tot}}{M_0}, \tag{163}$$

where the mass and radius scales are set by

$$\nu_c^2(0) = \frac{GM_0}{R_0},\ M_0 = 4\pi R_0^3 \rho_c(0), \tag{164}$$

so x and g are density contrasts, the ratios of the core density and mean halo density, respectively, to the critical density of the universe at the time of formation. Both are believed to be redshift-independent; we take $x = 1.8 \times 10^4$ (which gives $g \approx 206$), the value derived for truncated isothermal spheres formed from collapse top-hat density perturbations [254, 255]. Then $\rho_c(0) = x\rho_{\rm crit}(z_0)$, and

$$\nu_c^2(0) = GM_0\left(\frac{4\pi\rho_c(0)}{M_0}\right)^{1/3} = G\left(4\pi x\rho_{\rm crit}(z_0)\right)^{1/3}\left(\frac{M_{\rm tot}}{g(x)}\right)^{2/3}, \tag{165}$$

so

$$t_r(0) = (a\sigma f)^{-1}\left(x\rho_{\rm crit}(0)\right)^{-1} G^{-1/2}\left(4\pi x\rho_{\rm crit}(z_0)\right)^{-1/6}\left(\frac{M_{\rm tot}}{g(x)}\right)^{-1/3}, \tag{166}$$

and the arbitrary scales M_0, R_0 have canceled out of the expression, as desired. Hence (159)

becomes

$$t(z_1) - t(z_0) \geq \frac{490}{(4\pi)^{1/6}\sqrt{G}a\sigma f}\left(x\rho_{\rm crit}(z_0)\right)^{-7/6}\left(\frac{M_{\rm tot}}{g(x)}\right)^{-1/3}, \tag{167}$$

which corresponds to a lower bound on σ,

$$\sigma \geq \frac{490}{af\sqrt{G}}\frac{\left(x\rho_{\rm crit}(z_0)\right)^{-7/6}}{(4\pi)^{1/6}\left(t(z_1) - t(z_0)\right)}\left(\frac{M_{\rm tot}}{g(x)}\right)^{-1/3}. \tag{168}$$

Equation (159) was derived on the assumption that the SIDM halo passes through a self-similar stage on the way to core collapse. As we saw in subsection 6.4.5 above, however, self-similarity is broken when the core enters the optically thick regime, when $\lambda/H \ll 1$. If the cross section is sufficiently large that the core is already optically thick for the initial conditions shown in Figure 10, i.e. when $\tilde{\rho}^{int} = 1$, then this assumption will be violated: an NFW halo will *never* become self-similar. In this case, we are unable to use the collapse time and seed black hole mass estimates calculated in section 6. The cusp may collapse directly into a black hole, or, as Balberg and Shapiro [17] claim, strong shocks may entirely destroy any central structure: further work, beyond the scope of this paper, is needed. Regardless, we can use the condition that the self-similar core not initially be in the short mean free path regime to place an upper bound on the value of σ for which "standard" gravothermal collapse is possible:

$$\frac{\lambda}{H} = (f\rho\sigma)^{-1}/\sqrt{\nu^2\left(4\pi G\rho\right)^{-1}} = \frac{(4\pi)^{1/3}}{\sigma f}\left(x\rho_{\rm crit}(z_0)\right)^{-2/3}\left(\frac{M_{\rm tot}}{g(x)}\right)^{-1/3} \leq 1, \tag{169}$$

so

$$\sigma \leq \left(\frac{4\pi g(x)}{M_{\rm tot}}\right)^{1/3}\left(x\rho_{\rm crit}(z_0)\right)^{-2/3} f^{-1}. \tag{170}$$

Given the bounds (168,170), we can calculate the allowed range of σ given a halo mass $M_{\rm tot}$, SIDM mass fraction f, and collapse time z_1. Figure 14 plots this range as a function of redshift of formation for $M_{\rm tot} = 10^{11}M_\odot$, $f = 0.01$, and $z_1 = 7.9$, which we have chosen because

$$t(6) \approx t(7.9) + \ln 10^3 \times 45.1 \text{ Myr}, \tag{171}$$

i.e. because seed black holes formed at this redshift could grow three orders of magnitude by Eddington-limited gas accretion, as described in subsection 5.1, by $z = 6$.

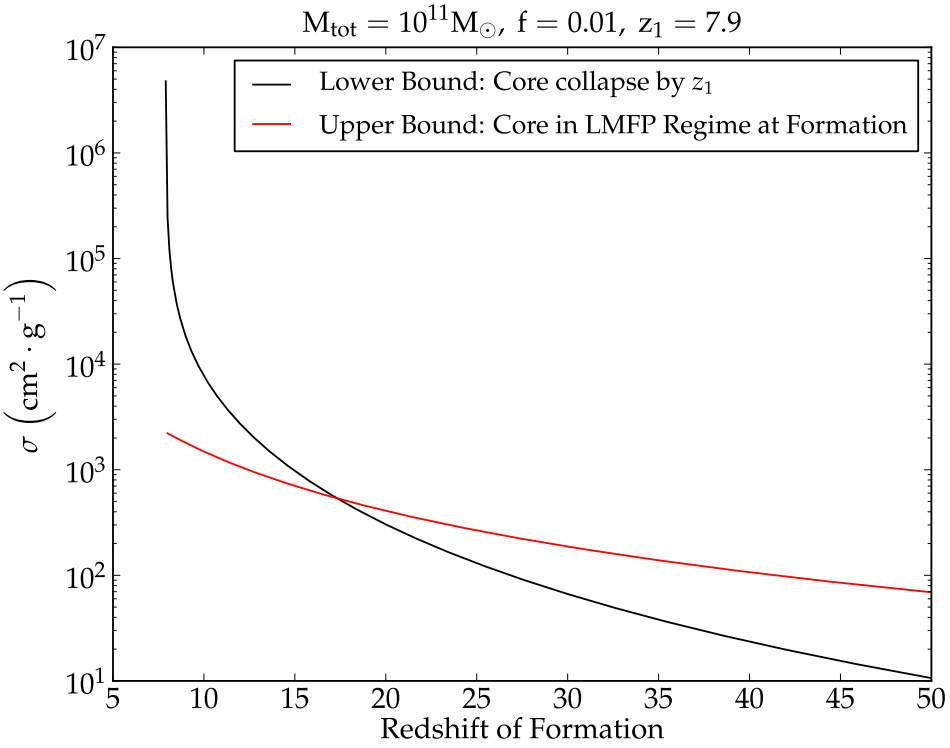

Figure 14: Upper and lower bounds on σ as a function of the redshift of formation z_0 for a $M_{\rm tot} = 10^{11} M_\odot$ halo with $f = 0.01$ collapsing at $z_1 = 7.9$, with enough time for a black hole to grow by a factor of 10^3 by redshift 6 from Eddington-limited gas accretion. As discussed in the text, the latest possible redshift of formation, $z_{0,\rm min} \sim 17.3$, where the upper and lower bounds cross, is independent of halo mass and SIDM fraction. For this choice of $M_{\rm tot}$ and f, the required cross section is $\sigma_{\rm max} \approx 530$ cm$^2 \cdot$ g^{-1}.

Note that the upper and lower bounds on σ (168,170) have identical dependences on $M_{\rm tot}$ and f. When we solve for the latest possible halo collapse time by setting the bounds equal to each other, then, $M_{\rm tot}$ and f cancel from the equation: $z_{0,\rm min}$ is independent of the halo parameters. Hence *all* halos that grow seed black holes by gravothermal collapse (at least the "normal" version where the core is not initially optically thick) must have formed by $z_{0,\rm min}$. The largest allowed cross section (i.e. where the core has an initial optical depth of unity) which leads to normal gravothermal collapse, starting at $z_{0,\rm min}$ and ending at z_1, *does* depend on the halo parameters: we see that it goes as f^{-1} and $M_{\rm tot}^{-1/3}$. For the choice of halo parameters in Figure 14, the largest allowed cross section is $\sigma_{\rm max} \approx 530\ {\rm cm}^2 \cdot {\rm g}^{-1}$. Smaller cross sections can produce a seed black hole of the same size, for values of σ that saturate the lower bound at earlier redshifts, but at the expense of requiring rarer halos, since the halo mass function is strongly downward-sloping at high redshifts, as discussed below.

7.2 Halo Abundance

In order to match the observed number density of $z \gtrsim 6$ halos, 1 ${\rm Gpc}^{-3}$ [192], we need to calculate the abundance of halos of a given mass at a given redshift, i.e. the halo mass function $\frac{dN}{dM}(z)$. We use the Sheth-Tormen analytical approximation [256], corrected to match the numerical results of the Bolshoi cosmological ΛCDM simulation [257]. This should be reasonable to use even in a cosmology containing SIDM as long as the cross sections considered do not impact halo formation, as determined by Table 1 on page 34. The corrected ST approximation to the halo mass function is given by

$$M\frac{dN}{dM} = \Omega_{M,0}\rho_{\rm cr,0}\frac{d\sigma(M)}{\sigma(M)dM}f(\sigma)F(\delta). \tag{172}$$

Here $\sigma(M)$ is the RMS density fluctuation, which can be approximated with better than 2% accuracy for masses larger than $10^7 h^{-1} M_\odot$ for the WMAP7 cosmological parameters [1] used in the Bolshoi simulation as

$$\sigma(M) = \frac{16.9y^{0.41}\delta(a)}{1 + 1.102y^{0.20} + 6.22y^{0.333}},\ y \equiv \frac{10^{12}h^{-1}M_\odot}{M}. \tag{173}$$

The (normalized) linear growth-rate function $\delta(a)$ is defined by

$$\delta(a) \equiv \frac{D(a)}{D(1)},\ a = \frac{1}{1+z}, \tag{174}$$

where $D(a)$ is the growth-rate factor, approximated to better than .1% accuracy at times when the energy density in radiation is negligible by [258, 259]

$$D(a) = \frac{(5/2)a\Omega_M(a)}{\Omega_M(a)^{4/7} - \Omega_\Lambda(a) + (1+\Omega_M(a)/2)(1+\Omega_\Lambda(a)/70)}, \tag{175}$$

with

$$\Omega_M(a) = \left(1 + \frac{\Omega_{\Lambda,0}}{\Omega_{M,0}}a^3\right)^{-1},\ \Omega_\Lambda(a) = 1 - \Omega_M(a). \tag{176}$$

The exponential suppression $f(\sigma)$ is given by

$$f(\sigma) = A\sqrt{\frac{2b}{\pi}}\left[1 + \left(bx^2\right)^{-0.3}\right] x \exp\left(-\frac{bx^2}{2}\right), \tag{177}$$

with $A = 0.322$, $b = 0.707$, and $x = 1.686/\sigma(M)$. Finally, the correction factor to match the Bolshoi results is

$$F(\delta) = \frac{(5.501\delta)^4}{1 + (5.500\delta)^4}. \tag{178}$$

Figure 15 plots the halo mass function at $z_0 = 17.3$. Evidently the function is extremely steep near 1 Gpc^{-3}, the number density of $z \geq 6$ SMBHs [192]: changing the halo mass by an order of magnitude, from 10^{10} to $10^{11}M_\odot$, alters the abundance by a factor of $\sim 10^5$. We see that $10^{11}M_\odot$ halos are too rare at this magnitude to explain the oberved abundance of high-redshift quasars: evidently the example in the last subsection is unsuccessful.

However, we now have a procedure for determining values of ($M_{\rm tot}$, f, σ) which *do* succeed in giving the correct quasar abundance, given only a choice of the amount of growth that comes from Eddington-limited gas accretion (or, alternatively, the size of the seed black holes grown by gravothermal collapse). Once this has been chosen, we find the necessary value of z_1 from an expression analogous to equation (171). We find the latest possible collapse redshift z_0 by setting the upper and lower bounds on σ (168,170) equal. We then calculate the halo mass function (172) at z_0, and find the mass $M_{\rm tot}$ which has a number density of 1 Gpc^{-3} at this redshit. Since we know how big we want the seed black holes to

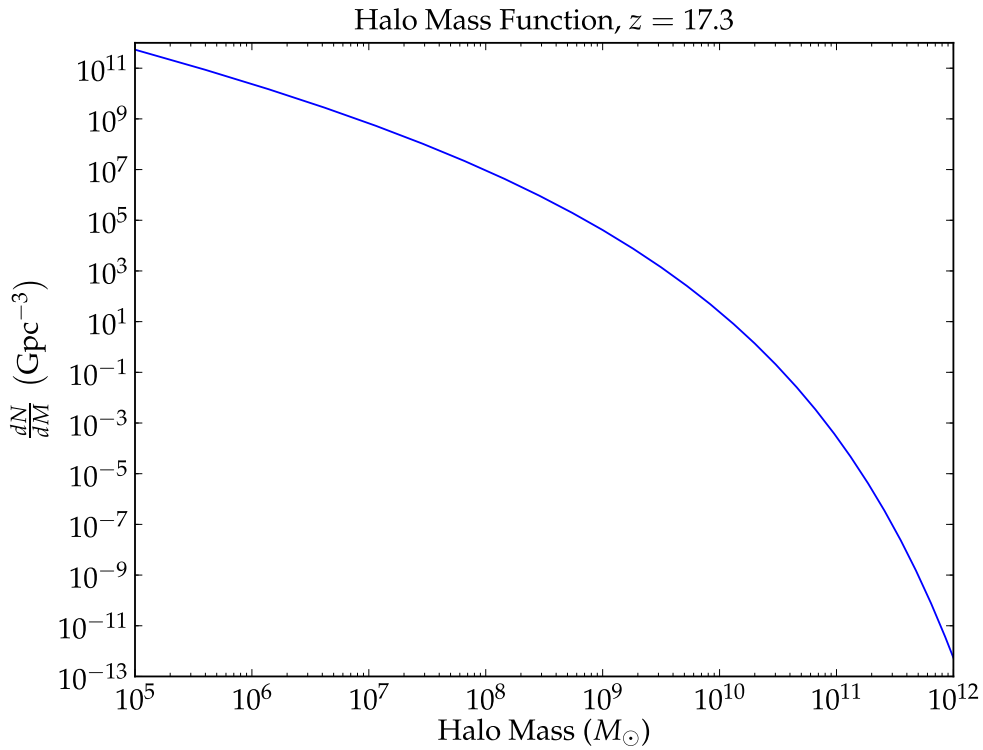

Figure 15: The halo mass function at $z_0 = 17.3$ in Gpc^{-3}. The function is very steep near $\frac{dN}{dM} \approx 1$ Gpc^{-3}: $10^{10} M_\odot$ halos have an abundance of 24.2 per cubic gigaparsec, while $10^{11} M_\odot$ halos have an abundance of only 2.9×10^{-4} Gpc^{-3}. halos with mass $\sim 2.15 \times 10^{10} M_\odot$ have the correct abundance required to explain the observed density of black holes at redshift 6.

$\frac{M_{\rm SMBH}}{M_{\rm seed}}$	$M_{\rm seed}$	z_1	z_0	$M_{\rm tot}(M_\odot)$	f	$\sigma_{\rm max}(\frac{\rm cm^2}{\rm g})$	$\sigma f(\frac{\rm cm^2}{\rm g})$
10^0	$10^9 M_\odot$	6	13.43	1.55×10^{11}	2.59	N/A	N/A
10^1	$10^8 M_\odot$	6.51	14.48	8.80×10^{10}	0.454	17.2	7.81
10^2	$10^7 M_\odot$	7.13	15.76	4.56×10^{10}	0.088	94.4	8.31
10^3	$10^6 M_\odot$	7.89	17.33	2.12×10^{10}	0.019	471	8.95
10^4	$10^5 M_\odot$	8.87	19.33	8.45×10^9	0.0047	$2.11 \cdot 10^3$	9.92
10^5	$10^4 M_\odot$	10.15	21.98	2.74×10^9	0.0015	$7.52 \cdot 10^3$	11.28
10^6	$10^3 M_\odot$	11.96	25.70	6.58×10^8	$6.08 \cdot 10^{-4}$	$2.21 \cdot 10^4$	13.44
10^7	$10^2 M_\odot$	14.74	31.43	9.66×10^7	$4.14 \cdot 10^{-4}$	$4.18 \cdot 10^4$	17.31
10^8	$10^1 M_\odot$	19.74	41.71	5.74×10^6	$6.96 \cdot 10^{-4}$	$3.67 \cdot 10^4$	25.54
10^9	$10^0 M_\odot$	32.43	67.85	3.49×10^4	0.011	$4.90 \cdot 10^3$	53.90

Table 2: Values of $M_{\rm tot}$, f, and σ that successfully yield the current high-redshift quasar abundance, given an amount of growth from Eddington-limited gas accretion, $M_{\rm SMBH}/M_{\rm seed}$, or the mass of the BH seeded by collapse, $M_{\rm seed}$. The first row shows that some gas accretion is still necessary: to get $M_{\rm seed} = 10^9 M_\odot$, an aphysical value of $f > 1$ is required. The last column makes clear that σf increases slowly as the seed BH mass decreases. Although smaller seed black holes are possible, there are not enough Salpeter e-folding times to grow the seeds to $10^9 M_\odot$ by $z \sim 6$: see the last paragraph of subsection 5.1. All of the successful entries on this table satisfy the constraints for valid NFW halo formation from Table 1. Some of the entires may violate the black hole accretion constraint (61), depending on the value of α assumed, but see the caveats regarding this constraint on page 37. Finally, some of the entries may violate the Yoshida *et al.* cluster core constraint, presented in subsection 4.2.1 on page 38: certainly the values that produce 10^8 and $10^7 M_\odot$ seed black holes do.

be, the choice of $M_{\rm tot}$ sets the SIDM mass density f via equation (157). Finally, we plug the values of f and $M_{\rm tot}$ into the bounds on σ to find the required value for collapse at z_1 from formation at z_0. Table 2 presents these values and comments on the regimes in which they avoid existing bounds. Each row of the table that evades the bounds constitutes a successful model of SMBH formation via gravothermal collapse.

8 Conclusion

In this paper, we have presented an alternative model of high-redshift supermassive black hole formation and evolution, in which seed black holes are grown by the gravothermal collapse of fractional self-interacting dark matter halos. In assembling the model, we drew on data, mechanisms, and simulations covering a vast hierarchy of scales: from microphysical motivations for a dark matter self-interacting cross section to Bondi accretion and core collapse on kiloparsec scales, galaxy formation and evolution on megaparsec scales, structure formation, the halo mass function, and observations of quasars and clusters on gigaparsec scales, and ultimately to the cosmological parameters observed in the cosmic microwave background, which bring us back full circle to the underlying microphysics. It is a testament to the vast amount of theoretical and experimental knowledge we have about the universe on all these scales that a model of this nature can actually be constrained severely enough to give specific predictions based on concrete scenarios; cosmological speculations of this sort are presently at the fruitful boundary between completely understood and entirely unknown.

References

[1] E. Komatsu *et al.*, *Astrophys. J. Suppl.* **192**, 18 (2011).

[2] V. Springel *et al.*, *Nature* **435**, 629 (2005).

[3] A. A. Klypin, S. Trijillo-Gomez, and J. Primack, *Astrophys. J.* **740**, 102 (2011).

[4] A. A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada, *Astrophys. J.* **522**, 82 (1999).

[5] B. Moore, S. Ghigna, F. Governato, G. Lake, T. Quinn, J. Stadel, and P. Tozzi, *Astrophys. J.* **524**, L19 (1999).

[6] M. Boylan-Kolchin, J. S. Bullock, and M. Kaplinghat, *Mon. Not. R. Astron. Soc.* **422**, 1203 (2012).

[7] B. Moore, *Nature* **370**, 629 (1994).

[8] B. Moore, F. Governato, T. Quinn, J. Stadel, and G. Lake, *Astrophys. J.* **499**, L5 (1998).

[9] E. Aprile *et al.* (XENON100), *Phys. Rev. Lett.* **107**, 131302 (2011).

[10] D. N. Spergel and P. J. Steinhardt, *Phys. Rev. Lett.* **84**, 3760 (2000).

[11] M. Markevitch *et al.*, *Astrophys. J.* **606**, 819 (2004).

[12] N. Yoshida, V. Springel, S. D. M. White, and G. Tormen, *Astrophys. J.* **544**, L87 (2000).

[13] J. F. Hennawi and J. P. Ostriker, *Astrophys. J.* **572**, 41 (2002)

[14] L. Hui, *Phys. Rev. Lett.* **86**, 3467 (2001).

[15] A. Loeb and N. Weiner, *Phys. Rev. Lett.* **106**, 171302 (2011).

[16] D. Lynden-Bell and R. Wood, *Mon. Not. R. Astron. Soc.* **138**, 495 (1968).

[17] S. Balberg and S. L. Shapiro, *Phys. Rev. Lett.* **88**, 101301 (2002).

[18] X. Fan *et al.*, *Astron. J.* **122**, 2833 (2001).

[19] X. Fan *et al.*, *Astron. J.* **125**, 1649 (2003).

[20] C. J. Willott *et al.*, *Astron. J.* **134**, 2435 (2007).

[21] Y. Li *et al.*, *Astrophys. J.* **665**, 187 (2007).

[22] D. J. Mortlock *et al.*, *Nature* **474**, 616 (2011).

[23] S. Balberg, S. L. Shapiro, and S. Inagaki, *Astrophys. J.* **568**, 475 (2002).

[24] J. Koda and P. R. Shapiro, *Mon. Not. R. Astron. Soc.* **415**, 1125 (2011).

[25] J. Dunkley *et al.*, *Astrophys. J.* **739**, 52 (2011).

[26] D. J. Fixsen *et al.*, *Astrophys. J.* **473**, 576 (1996).

[27] J. C. Mather *et al.*, *Astrophys. J.* **512**, 511 (1999).

[28] D. J. Fixsen, *Astrophys. J.* **707**, 916 (2009).

[29] S. Weinberg, *Cosmology* (Oxford University Press, Oxford, England, 2008), section 7.2.

[30] G. Steigman, *Annu. Rev. Nucl. Part. S.* **57**, 463 (2007)

[31] G. Steigman, *J. Cosmol. Astropart. P.* **4,** 29 (2010).

[32] A. G. Reiss *et al.*, *Astrophys. J.* **699**, 539 (2009).

[33] A. G. Reiss *et al.*, *Astrophys. J.* **730**, 119 (2011).

[34] M. Hicken *et al.*, *Astrophys. J.* **700**, 1097 (2009).

[35] R. Kessler *et al.*, *Astrophys. J. Suppl.* **185**, 32 (2009).

[36] L. Anderson *et al.*, arXiv:1203.6594 (2012).

[37] A. G. Sanchez *et al.*, arXiv:1203.6619 (2012).

[38] A. Vikhlinin *et al.*, *Astrophys. J.* **692**, 1060 (2009).

[39] A. Mantz, S. W. Allen, D. Rapetti, and H. Ebeling, *Mon. Not. R. Astron. Soc.* **406**, 1759 (2010).

[40] R. Fadely, C. R. Keeton, R. Nakajima, and G. M. Bernstein, *Astrophys. J.* **711**, 246 (2010).

[41] E. Jullo *et al.*, *Science* **329**, 924 (2010).

[42] T. Schrabback *et al.*, *Astron. Astrophys.* **516**, A63 (2010).

[43] E. Semboloni *et al.*, *Mon. Not. R. Astron. Soc.* **410**, 143 (2011).

[44] G. R. Blumenthal, H. Pagels, and J. R. Primack, *Nature* **299**, 37 (1982).

[45] M. Kamionkowski and A. R. Liddle, *Phys. Rev. Lett.* **84**, 4525 (1999).

[46] V. K. Narayanan, D. N. Spergel, R. Davé, and C.-P. Ma, *Astrophys. J.* **543**, L103 (2000).

[47] S. Weinberg, *Cosmology* (Oxford University Press, Oxford, England, 2008), section 3.4.

[48] G. Jungman, M. Kamionkowski, and K. Griest, *Phys. Rep.* **267**, 195 (1996).

[49] K. Nakamura *et al.*, (Particle Data Group), *J. Phys.* **G37**, 075021 (2010).

[50] M. W. Goodman and E. Witten, *Phys. Rev.* **D31**, 3059 (1985).

[51] A. Bottino, F. Donato, N. Fornengo, and S. Scopel, *Phys. Rev.* **D72**, 83521 (2005).

[52] R. Catena and P. Ullio, *J. Cosmol. Astropart. P.* **8**, 4 (2010).

[53] M. C. Smith *et al.*, *Mon. Not. R. Astron. Soc.* **379**, 755 (2007).

[54] A. K. Drukier, K. Freese, and D. N. Spergel, *Phys. Rev.* **D33**, 3495 (1986).

[55] R. Bernabei *et al.* (DAMA/LIBRA), *Eur. Phys. J.* **C67**, 39 (2010).

[56] C. E. Aalset *et al.* (CoGeNT), *Phys. Rev. Lett.* **107**, 141301 (2011).

[57] Z. Ahmed *et al.* (CDMS), arXiv:1203.1309 (2012).

[58] E. Aprile *et al.* (XENON100), *Phys. Rev. Lett.* **10**5, 131302 (2010).

[59] E. Armengaud *et al.* (EDELWEISS), *Phys. Lett.* **B702**, 329 (2011).

[60] Z. Ahmed *et al.* (CDMS), *Science* **327**, 1619 (2010).

[61] Z. Ahmed *et al.* (CDMS), *Phys. Rev. Lett.* **106**, 131302 (2011).

[62] J. Angle *et al.* (XENON10), *Phys. Rev. Lett.* **107**, 51301 (2011).

[63] C. E. Aalseth *et al.* (CoGeNT), *Phys. Rev. Lett.* **106**, 131301 (2011).

[64] C. Savage *et al.*, *J. Cosmol. Astropart. P.* **04**, 10 (2009).

[65] R. Trotta *et al.*, *J. High Energy Phys.* **12**, 24 (2008).

[66] O. Buchmueller *et al.*, *Eur. Phys. J.* **C71**, 1634 (2011).

[67] M. Drees and G. Gerbier, *Dark Matter*, in K. Nakamura *et al.*, (Particle Data Group), *J. Phys.* **G37**, 075021 (2010) (September 2011 update for the 2012 edition).

[68] N. Katz and S. D. M. White, *Astrophys. J.* **412**, 455 (1993).

[69] F. J. Summers, M. Davis, and A. E. Evrard, *Astrophys. J.* **454**, 1 (1995).

[70] B. Moore, N. Katz, and G. Lake, *Astrophys. J.* **457**, 455 (1996).

[71] B. Moore, F. Governato, T. Quinn, G. Lake, and J. Stadel, *Astrophys. J.* **499**, L5 (1998).

[72] S. Ghigna, B. Moore, F. Governato, G. Lake, T. Quinn, and J. Stadel, *Mon. Not. R. Astron. Soc.* **300**, 146 (1998)

[73] A. A. Klypin, S. Gottloeber, A. V. Kravtsov, and M. Khokhlov, *Astrophys. J.* **516**, 530 (1998).

[74] M. Mateo, *Ann. Rev. Astron. Astr.* **36**, 435 (1998).

[75] B. Binggeli, A. Sandage, and G. A. Tammann, *Astron. J.* **90**, 1681 (1985).

[76] B. Willman *et al.*, *Astron. J.* **129**, 2692 (2005).

[77] E. J. Tollerud, J. S. Bullock, L. E. Strigari, and B. Willman, *Astrophys. J.* **688**, 277 (2008).

[78] J. Diemand, M. Kuhlen, P. Madau, M. Zemp, B. Moore, D. Potter, and J. Stadel, *Nature* **454**, 735 (2008).

[79] V. Springel *et al.*, *Mon. Not. R. Astron. Soc.* **391**, 1685 (2008).

[80] F. Stoehr, S. D. M. White, G. Tormen, and V. Springel, *Mon. Not. R. Astron. Soc.* **335**, L84 (2002)

[81] E. Hayashi, J. F. Navarro, J. E. Taylor, J. Stadel, and T. Quinn, *Astrophys. J.* **584**, 541 (2003).

[82] J. Peñarrubia, A. W. McConnachie, and J. F. Navarro, *Astrophys. J.* **672**, 904 (2008).

[83] J. S. Bullock, A. V. Kravtsov, and D. H. Weinberg, *Astrophys. J.* **539**, 517 (2000).

[84] A. V. Kravtsov, O. Y. Gnedin, and A. K. Klypin, *Astrophys. J.* **609**, 482 (2004).

[85] M. Ricotti and N. Y. Gnedin, *Astrophys. J.* **629**, 259 (2005).

[86] J. Dubinski and R. G. Carlberg, *Astrophys. J.* **378**, 496 (1991).

[87] R. A. Flores and J. R. Primack, *Astrophys. J.* **427**, L1 (1994).

[88] J. F. Navarro, C. S. Frenk, and S. D. M. White, *Astrophys. J.* **490**, 493 (1997).

[89] S. Ghigna, B. Moore, F. Governato, G. Lake, T. Quinn, and J. Stadel, *Astrophys. J.* **544**, 616 (2000).

[90] D. Merritt, J. F. Navarro, A. Ludlow, and A. Jenkins, *Astrophys. J.* **624**, L85 (2005).

[91] A. W. Graham, D. Merritt, B. Moore, J. Diemand, and B. Terzić, *Astron. J.* **132**, 2701 (2006).

[92] J. F. Navarro *et al.*, *Mon. Not. R. Astron. Soc.* **402**, 21 (2010).

[93] J. Stadel *et al.*, *Mon. Not. R. Astron. Soc.* **398**, L21 (2009).

[94] S.-H. Oh, W. J. G. de Blok, E. Brinks, F. Walter, and R. C. Kennicut, *Astron. J.* **141**, 193 (2011).

[95] F. C. van den Bosch, B. E. Robertson, J. J. Dalcanton, and W. J. G. de Blok, *Astron. J.* **119**, 579 (2000).

[96] R. A. Swaters, B. F. Madore, F. C. van den Bosch, and M. Balcells, *Astrophys. J.* **583**, 732 (2003).

[97] G. Rhee, O. Valenzuela, A. Klypin, J. Holtman, and B. Moorthy, *Astrophys. J.* **617**, 1059 (2004).

[98] G. R. Blumenthal, S. M. Faber, R. Flores, and J. R. Primack, *Astrophys. J.* **301**, 27 (1986).

[99] J. F. Navarro, C. S. Frenk, and S. D. M. White, *Astrophys. J.* **462**, 563 (1996).

[100] J. I. Read and G. Gilmore, *Mon. Not. R. Astron. Soc.* **356**, 107 (2005).

[101] O. Y. Gnedin and H. Zhao, *Mon. Not. R. Astron. Soc.* **333**, 299 (2002).

[102] A. El-Zant, I. Shlosman, and Y. Hoffman, *Astrophys. J.* **560**, 636 (2001).

[103] F. Governato *et al.*, *Nature* **463**, 203 (2010).

[104] J. Binney, O. Gerhard, and J. Silk, *Mon. Not. R. Astron. Soc.* **321**, 471 (2001).

[105] S. Mashcheko, H. M. P. Couchman, and J. Wadsley, *Nature* **442**, 539 (2006).

[106] H. J. Mo and S. Mao, *Mon. Not. R. Astron. Soc.* **353**, 829 (2004).

[107] S.-H. Oh *et al.*, *Astron. J.* **142**, 24 (2011).

[108] G. L. Bryan and M. L. Norman, *Astrophys. J.* **495**, 80 (1998).

[109] T. D. Lee and Y. Yang, *Phys. Rev.* **104**, 254 (1956).

[110] E. W. Kolb, D. Seckel, and M. Turner, *Nature* **514**, 415 (1985).

[111] R. Foot, H. Lew, and R. Volkas, *Phys. Lett.* **B272**, 67 (1991).

[112] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hopes, and R. R. Hudson, *Phys. Rev.* **105**, 1413 (1957).

[113] R. L. Garwin, L. M. Lederman, and M. Weinrich, *Phys Rev.* **105**, 1415 (1957).

[114] Z. Berezhiani and R. N. Mohapatra, *Phys. Rev.* **D52**, 6607 (1995).

[115] Z. Berezhiani, A. D. Dolgov, and R. N. Mohapatra, *Phys. Lett..* **B**375, 26 (1996).

[116] R. N. Mohapatra and V. L. Teplitz, *Phys. Rev.* **D62**, 63506 (2000).

[117] P. J. Steinhardt, private communication (2012).

[118] A. Kusenko and P. J. Steinhardt, *Phys. Rev. Lett.* **87**, 141301 (2001).

[119] G. Rosen, *J. Math. Phys.* **9**, 996 (1968).

[120] R. Friedberg, T. D. Lee, and A. Sirlin, *Phys. Rev.* **D13**, 2739 (1976).

[121] S. Coleman, *Nucl. Phys.* **B262**, 263 (1985).

[122] T. D. Lee and Y. Pang, *Phys. Rept.* **221**, 251 (1992).

[123] J. A. Frieman, G. B. Gelmini, M. Gleiser, and E. W. Kolb, *Phys. Rev. Lett.* **60**, 2101 (1988).

[124] K. Griest and E. W. Kolb, *Phys. Rev.* **D40**, 3231 (1989).

[125] J. A. Frieman, A. V. Olinto, M. Gleiser, and C. Alcock, *Phys. Rev.* **D40**, 3241 (1989).

[126] I. Affleck and M. Dine, *Nucl. Phys.* **B249**, 361 (1985).

[127] M. Dine, L. Randall, and S. Thomas, *Phys. Rev. Lett.* **75**, 398 (1995).

[128] M. Dine, L. Randall, and S. Thomas, *Nucl. Phys.* **B458**, 291 (1996).

[129] M. J. G. Veltman and F. J. Ynduráin, *Nucl. Phys.* **B325**, 1 (1989).

[130] V. Silveira and A. Zee, *Phys. Lett.* **B161**, 136 (1985).

[131] J. McDonald, *Phys. Rev.* **D50**, 3637 (1994).

[132] M. C. Bento, O. Bertolami, R. Rosenfeld, and L. Teodoro, *Phys. Rev.* **D62**, 41302 (2000).

[133] C. P. Burgess, M. Pospelov, and T. ter Veldhuis, *Nucl. Phys.* **B619**, 709 (2001).

[134] L. B. Okun, *Nucl. Phys.* **B173**, 1 (1980).

[135] A. E. Faraggi and M. Pospelov, *Astropart. P.* **16**, 451 (2002).

[136] A. E. Faraggi, D. V. Nanopoulos, and K. Yuan, *Nucl. Phys.* **B335**, 347 (1990).

[137] A. E. Faraggi, *Phys. Lett.* **B278**, 131 (1992).

[138] M. Visser, *Phys. Lett.* **B159**, 22 (1985).

[139] L. Randall and R. Sundrum, *Phys.Rev. Lett.* **83**, 3370 (1999).

[140] M. Gogberashvili, *Europhys. Lett.* **49**, 396 (2000).

[141] P. Binétruy, C. Deffayet, and D. Langlois, *Nucl. Phys.* **B565**, 269 (2000).

[142] E. J. Copeland, A. R. Liddle, and J. E. Lidsey, *Phys. Rev.* **D64**, 23509 (2001).

[143] R. A. Frewin and J. E. Lidsey, *Int. J. Mod. Phys.* **D2**, 323 (1993).

[144] S. Barshay and G. Kreyerhoff, *Eur. Phys. J.* **C5**, 369 (1998).

[145] T. Matos and L. A. Ureña-López, *Class. Quantum Grav.* **17**, L75 (2000).

[146] T. Matos and L. A. Ureña-López, *Phys. Lett.* **B538**, 246 (2002).

[147] J. E. Lidsey, T. Matos, and L. A. Ureña-López, *Phys. Rev.* **D66**, 23514 (2002).

[148] S. W. Hawking, *Nature* **248**, 30 (1974).

[149] J. D. Barrow, E. J. Copeland, E. W. Kolb, and A. R. Liddle, *Phys. Rev.* **D43**, 984 (1991).

[150] M. Dine, W. Fischler, and M. Srednicki, *Nucl. Phys.* **B189**, 575 (1981).

[151] S. Dimopoulos and S. Raby, *Nucl. Phys.* **B192**, 353 (1981).

[152] J. L. Feng, H. Tu, and H.-B. Yu, *J. Cosmol. Astropart. P.* **10**, 43 (2008).

[153] J. L. Feng and J. Kumar, *Phys. Rev. Lett.* **101**, 231301 (2008).

[154] J. L. Feng, M. Kaplinghat, H. Tu, and H.-B. Yu, *J. Cosmol. Astropart. P.* **7**, 4 (2009).

[155] N. Arkani-Hamed, D. P. Finkbeiner, T. R. Slatyer, and N. Weiner, *Phys. Rev.* **D79**, 15014 (2009).

[156] O. Adriani *et al.* (PAMELA), *Nature* **458**, 607 (2009).

[157] J. Chang *et al.*, *Nature* **456**, 362 (2008).

[158] A. W. Strong *et al.*, *Astron. Astrophys.* **444**, 495 (2005).

[159] D. J. Thompson, D. L. Bertsch, and R. H. O'Neal, Jr, *Astrophys. J. Suppl. Ser.* **157**, 324 (2005).

[160] A. Sommerfeld, *Ann. Phys. (Leipzig)* **403**, 257 (1931).

[161] J. L. Feng, M. Kaplinghat, and H.-B. Yu, *Phys. Rev. Lett.* **104**, 151301 (2010).

[162] M. R. Buckley and P. J. Fox, *Phys. Rev.* **D81**, 83522 (2010).

[163] S. A. Khrapak, A. V. Ivlev, G. E. Morfill, and S.K. Zhdanov, *Phys. Rev. Lett.* **90**, 225002 (2003).

[164] M. Vogelsberger, J. Zavala, and A. Loeb, arXiv:1201.5892 (2012).

[165] B. Batell, M. Prospelov, and A. Ritz, *Phys. Rev.* **D79**, 115019 (2009).

[166] D. P. Finkbeiner and N. Weiner, *Phys. Rev.* **D7**6, 83519 (2007).

[167] P. W. Graham, R. Harnik, S. Rajendran, and P. Saraswat, *Phys. Rev.* **D82**, 63512 (2010).

[168] H. S. Zhao, *Mon. Not. R. Astron. Soc.* **278**, 488 (1996).

[169] V. R. Eke, J. F. Navarro, and M. Steinmetz, *Astrophys. J.* **554**, 114 (2001).

[170] http://www.astro.uvic.ca/~jfn/cens/.

[171] H. Bondi, *Mon. Not. R. Astron. Soc.* **112**, 195 (1952).

[172] K. Gebhardt *et al.*, *Astron. J.* **122**, 2469 (2001).

[173] M. S. Seigar, *ISRN Astron. Astrophys.* 725697 (2011).

[174] C. S. Kockanek and M. White, *Astrophys. J.* **543**, 514 (2000).

[175] T. Kitayama and Y. Suto, *Astrophys. J.* **469**, 480 (1996).

[176] J. A. Tyson, G. P. Kochanski, and I. P. dell'Antonio, *Astrophys. J.* **498**, L107.

[177] J.S. Arabadjis *et al.*, *Astrophys. J.* **572**, 66 (2001).

[178] D. Clowe *et al., Astrophys. J.* **604**, 596 (2004).

[179] S.W. Randall *et al.*, *Astrophys. J.* **679**, 1173 (2008).

[180] O.Y. Gnedin and J.P. Ostriker, *Astrophys. J.* **561**, 61 (2001).

[181] J. S. Arabadjis, M. W. Bautz, and G. P. Garmire, *Astrophys. J.* **572**, 66 (2002).

[182] R. Genzel, A. Eckart, T. Ott, and F. Eisenhauer, *Mon. Not. R. Astron. Soc.* **291**, 219 (1997).

[183] A. M. Ghez, B. L. Klein, M. Morris, and E. E. Becklin, *Astrophys. J.* **509**, 678.

[184] J. Kormendy and D. Richstone, *Ann. Rev. Astron. Astr.* **33**, 581 (1995).

[185] S. Tremaine *et al.*, *Astrophys. J.* **574**, 740 (2002).

[186] D. Richstone *et al.*, *Nature* **385**, A14 (1998).

[187] Y. B. Zeldovich, *Dokl. Akad. Nauk. SSSR.* **155**, 67 (1964).

[188] E. E. Salpeter, *Astrophys. J.* **140**, 796 (1964).

[189] D. Lynden-Bell, *Nature* **223**, 690 (1969).

[190] M. Elvis *et al.*, *Astrophys. J. Suppl. Ser.* **95**, 1 (1994).

[191] Z. Shang *et al.*, *Astrophys. J. Suppl. Ser.* **196**, 2 (2011).

[192] Z. Haiman, arXiv:1203.6075 (2012).

[193] N. L. Shakura and R. A. Sunyaev, *Astron. Astrophys.* **24**, 337 (1973).

[194] J. P. Ostriker and N. Y. Gnedin, *Astrophys. J.* **472**, L63 (1996).

[195] M. Tegmark, J. Silk, M. J. Rees, A. Blanchard, T. Abel, and F. Palla, *Astrophys. J.* **474**, 1 (1997).

[196] V. Bromm, P. S. Coppin, and R. B. Larson, *Astrophys. J.* **527**, L5 (1999).

[197] T. Abel, G. L. Bryan, and M. L. Norman, *Astrophys. J.* **540**, 39 (2000).

[198] L. Gao, N. Yoshida, T. Abel, C. S. Frenk, A. Jenkins, and V. Springel, *Mon. Not. R. Astron. Soc.* **378**, 449 (2007).

[199] M. J. Turk, T. Abel, and B. O'Shea, *Science* **325**, 601 (2009).

[200] A. Stacy, T. H. Greif, and V. Bromm, *Mon. Not. R. Astron. Soc.* **403**, 405 (2010).

[201] C. F. McKee and J. C. Tan, *Astrophys. J.* **681**, 771 (2008).

[202] A. Heger, C. L. Fryer, S. E. Woosley, N. Langer, and D. H. Hartmann, *Astrophys. J.* **591**, 288 (2003).

[203] W. Zhang, S. E. Woosley, and A. Heger, *Astrophys. J.* **679**, 639 (2008).

[204] M. Volonteri, *Astron. Astrophys. Rev.* **18**, 279 (2010).

[205] C. L. Fryer, S. E. Woosley, and A. Heger, *Astrophys. J.* **550**, 372 (2001).

[206] J. R. Bond, W. D. Arnett, and B. J. Carr, *Astrophys. J.* **280**, 825 (1984).

[207] A. Loeb and F. A. Rasio, *Astrophys. J.* **432**, 52 (1994).

[208] D. J. Eisenstein and A. Loeb, *Astrophys. J.* **443**, 11 (1995).

[209] V. Bromm and A. Loeb, *Astrophys. J.* **596**, 34 (2003).

[210] M. Dijkstra, Z. Haiman, A. Mesinger, and J. S. B. Wyithe, *Mon Not. R. Astron. Soc.* **391**, 1961 (2008).

[211] M. Spaans and J. Silk, *Astrophys. J.* **652**, 902 (2006).

[212] M. C. Begeman and I. Shlosman, *Astrophys. J.* **702**, L5 (2009).

[213] H. J. Mo, S. Mao, and S. D. M. White, *Mon. Not. R. Astron. Soc.* **295**, 319 (1998).

[214] S. P. Oh and Z. Haiman, *Astrophys. J.* **569**, 558 (2002).

[215] S. M. Koushiappas, J. S. Bullock, and A. Dekel, *Mon. Not. R. Astron. Soc.* **354**, 292 (2004).

[216] I. Shlosman, J. Frank, and M. C. Begelman, *Nature* **338**, 45 (1989).

[217] G. Lodato and P. Natarajan, *Mon. Not. R. Astron. Soc.* **371**, 1813 (2006).

[218] J. H. Wise, M. J. Turk, and T. Abel, *Astrophys. J.* **682**, 745 (2008).

[219] M. C. Begelman and M. J. Rees, *Mon. Not. R. Astron. Soc.* **185**, 847 (1978).

[220] L. Spitzer, *Dynamical Evolution of Globular Clusters* (Princeton University Press, Princeton, NJ, 1987).

[221] D. C. Heggie, *Mon. Not. R. Astron. Soc.* **173**, 729 (1975).

[222] P. Hut *et al.*, *Publ. Astron. Soc. Pac.* **104**, 981 (1992).

[223] H. M. Lee, *Astrophys. J.* **319**, 801 (1987).

[224] G. D. Quinlan and S. L. Shapiro, *Astrophys. J.* **356**, 483 (1990).

[225] L. Spitzer, *Astrophys. J.* **158**, L139 (1969).

[226] E.T. Vishniac, *Astrophys. J.* **223**, 986 (1978).

[227] S. F. Portegies Zwart, J. Makino, S. L. W. McMillan, and P. Hut, *Astron. Astrophys.* **348**, 117 (1999).\

[228] S. Goswami, S. Umbreit, M. Bierbaum, and F. A. Rasio, arXiv:1105.5884 (2011).

[229] J. L. Johnson and V. Bromm, *Mon. Not. R. Astron. Soc.* **374**, 1557 (2007).

[230] T. Abel, J. H. Wise, and G. L. Byran, *Astrophys. J.* **659**, L87 (2007).

[231] N. Yoshida, S. P. Oh, T. Kitayama, and L. Hernquist, *Astrophys. J.* **663**, 687 (2007).

[232] J. M. Bardeen, W. H. Press, and S. A. Teukolsky, *Astrophys. J.* **178**, 347 (1972).

[233] S. N. Zhang, W. Cui, and W. Chen, *Astrophys. J.* **482**, L155 (1997).

[234] R. Narayan and J. E. McClintock, *Mon. Not. R. Astron. Soc.* **419**, L69 (2012).

[235] A. Escala, R. B. Larson, P. S. Coppi, and D. Mardones, *Astrophys. J.* **607**, 765 (2004).

[236] V. Springel, T. Di Matteo, and L. Hernquist, *Mon. Not. R. Astron. Soc.* **361**, 776 (2005).

[237] W. B. Bonnor and M. A. Rotenberg, *Proc. R. Soc. London Ser.-A* **265**, 109 (1961).

[238] A. Peres, *Phys. Rev.* **128**, 2471 (1962).

[239] M. J. Fitchett, *Mon. Not. R. Astron. Soc.* **203**, 2049 (1983).

[240] D. Merritt, M. Milosavljević, M. Favata, S. A. Hughes, and D. E. Holz, *Astrophys. J.* **607**, L9 (2004).

[241] Z. Haiman, *Astrophys. J.* **613**, 36 (2004).

[242] L. Blanchet, M. S. S. Qusailah, and C. M. Will, *Astrophys. J.* **635**, 508 (2005).

[243] J. G. Baker, J. Centrella, D.-I. Choi, J. R. van Meter, and M. C. Miller, *Astrophys. J.* **653**, L93 (2006).

[244] J. A. González, U. Sperhake, B. Brügmann, M. Hannam, and S. Husa, *Phys. Rev. Lett.* **98**, 91101 (2007).

[245] M. Campanelli, C. Lousto, Y. Zlochower, and D. Merritt, *Astrophys. J.* **659**, L5 (2007).

[246] J. A. González, M. Hannam, U. Sperhake, B. Brügmann, and S. Husa, *Phys. Rev. Lett.* **98**, 231101 (2007).

[247] T. Bogdanović, C. S. Reynolds, and M. C. Miller, *Astrophys. J.* **661**, L147 (2007).

[248] J. G. Baker *et al.*, *Astrophys. J.* **682**, L29 (2008).

[249] D. Lynden-Bell and P. P. Eggleton, *Mon. Not. R. Astron. Soc.* **191**, 483 (1980).

[250] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases* (Cambridge University Press, Cambridge, England, 1939).

[251] E. M. Lifshitz and L. P. Pitaevskii, *Physical Kinetics* (Oxford University Press, Oxford, England, 1981), chapter 1, problem 3.

[252] F. Reif, *Fundamentals of Statistical and Thermal Physics* (Waveland Press, Long Grove, IL, 2009), section 12.2.

[253] http://www.gnu.org/software/gsl/.

[254] P. R. Shapiro, I. T. Iliev, and A. C. Raga, *Mon. Not. R. Astron. Soc.* **307**, 203 (1999).

[255] I. T. Iliev and P. R. Shapiro, *Mon. Not. R. Astron. Soc.* **325**, 468.

[256] R. K. Sheth and G. Tormen, *Mon. Not. R. Astron. Soc.* **329**, 61 (2002).

[257] A. A. Klypin, S. Trujillo-Gomez, and J. Primack, *Astrophys. J.* **740**, 102 (2011).

[258] O. Lahav, P. B. Lilje, J. R. Primack, and M. J. Rees, *Mon. Not. R. Astron. Soc.* **251**, 128 (1991).

[259] S. M. Carroll, W. H. Press, and E. L. Turner, *Ann. Rev. Astron. Astrophys.* **30**, 499 (1992).